Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour

Tianyi Zhang
Singapore Management University
Singapore, Singapore
tianyizhang.2023@phdcs.smu.edu.sg

Colin Au Yeung
University of Calgary
Calgary, Canada
colin.auyeung@ucalgary.ca

Emily Aurelia
Singapore Management University
Singapore, Singapore
eaurelia.2021@scis.smu.edu.sg

Yuki Onishi
University of Bergen
Bergen, Norway
yuki.onishi@uib.no

Neil Chulpongsatorn
University of Calgary
Calgary, Canada
thobthai.chulpongsat@ucalgary.ca

Jiannan Li
Singapore Management University
Singapore, Singapore
jiannanli@smu.edu.sg

Anthony Tang
Singapore Management University
Singapore, Singapore
tonyt@smu.edu.sg

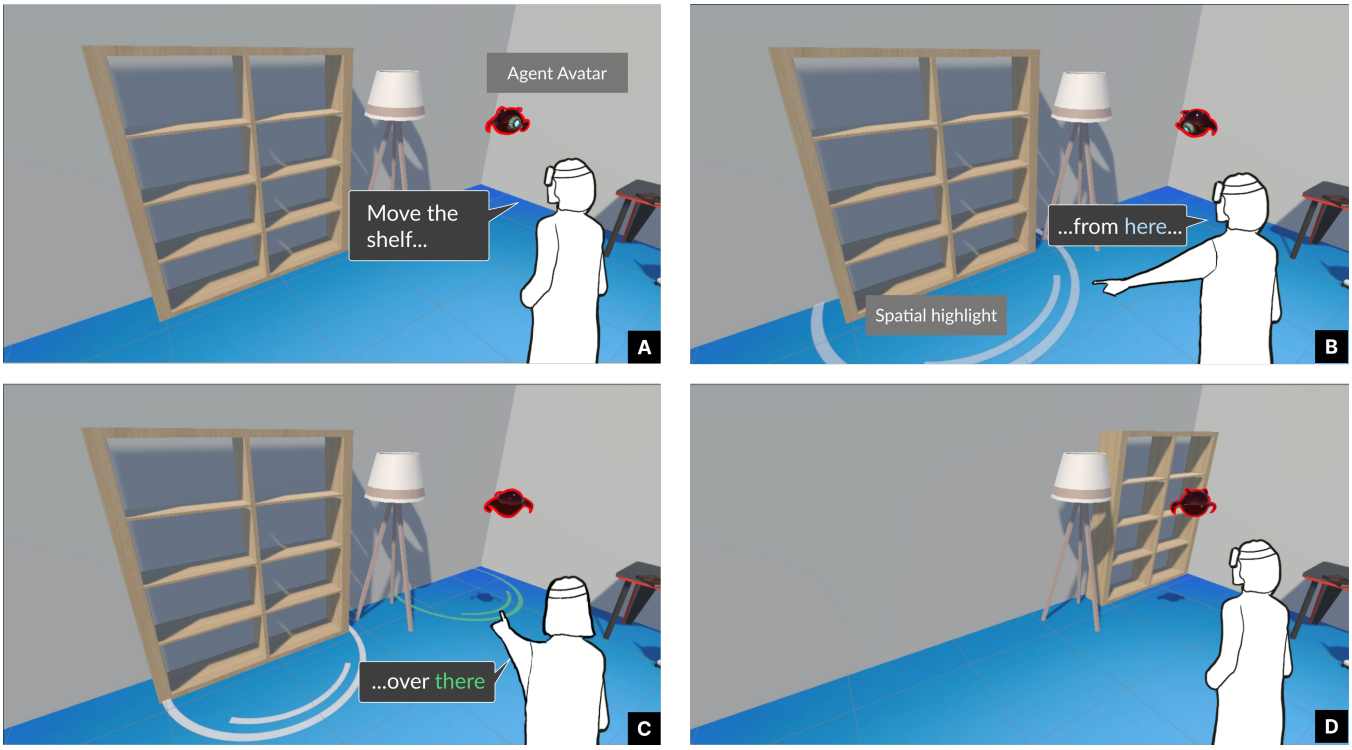


Figure 1: An AI agent's avatar can signal its understanding of the user's prompt by swivelling its orientation, as well as by creating spatial cues at the targets of the prompt. Here, the user asks to move a shelf. As she verbalizes the instruction, the avatar turns to look at the shelf to be moved, and then to the target location. Spatial cues can also highlight the shelf and where it would be moved to (B and C). We show these behaviors improve user satisfaction and help error prevention.

Abstract

Current voice agents wait for a user to complete their verbal instruction before responding; yet, this is misaligned with how humans engage in everyday conversational interaction, where interlocutors use multimodal signaling (e.g. nodding, grunting, or looking at referred to objects) to ensure conversational grounding. We designed an embodied VR agent that exhibits multimodal signaling behaviors in response to situated prompts, by turning its head, or by visually highlighting objects being discussed or referred to. We explore how people prompt this agent to design and manipulate the objects in a VR scene. Through a Wizard of Oz study, we found that participants interacting with an agent that indicated its understanding of spatial and action references were able to prevent errors 30% of the time, and were more satisfied and confident in the agent's abilities. These findings underscore the importance of designing multimodal signaling communication techniques for future embodied agents.

CCS Concepts

• **Human-centered computing** → *Interaction design*; **Empirical studies in interaction design**; **Natural language interfaces**; **Virtual reality**.

Keywords

situated prompting, multimodal signaling, common ground, human-ai collaboration

ACM Reference Format:

Tianyi Zhang, Colin Au Yeung, Emily Aurelia, Yuki Onishi, Neil Chulpongsatorn, Jiannan Li, and Anthony Tang. 2025. Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3706598.3713110

1 Introduction

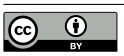
Intelligent voice agents, currently found in devices like phones and smart assistants, primarily rely on voice commands to execute tasks [63]. Yet, interactions with these agents are often limited to rigid, turn-based exchanges, where users issue voice commands, and agents respond verbally [37]. This can make the interaction feel overly formal and disconnected from the physical world.

However, communication in human interactions extends beyond verbal exchanges [17]. People often employ non-verbal cues while others are speaking—such as eye contact [57, 60], gestures [42], and subtle acknowledgments [17]—which help both speaker and listener to ensure they have a shared, mutual understanding of the conversation topic. These multimodal signals support local grounding, a form of common ground that refers to the shared understanding of the immediate physical context [16], such as the presence of specific objects and the spatial relationships between them. Local grounding is crucial in interactions involving the physical environment, where non-verbal cues—like a glance toward an object or a nod of acknowledgment—help participants confirm that they are aligned in their understanding of what is being referenced or acted upon [28].

Embodied agents can leverage a combination of communication modalities for signaling, including gaze, gestures, and spatial overlays through projection. Previous research with both screen-based virtual agents (e.g. [14, 15]) well as with social robotics have highlighted how gaze can function as a social mechanism to convey interest and understanding[45, 46] and how gestures or visual indicators can complement verbal communication [20, 31]. Building on these insights, this work explores the design of local grounding behaviors for embodied agents that respond through multimodal signals. Since such voice prompts are often *situated*—that is, referring to objects and actions within the immediate physical environment—it may be valuable for agents to possess a visible embodiment, which helps users ensure their instructions are understood properly. As illustrated in Figure 1, rather than verbal turns, our agents provide multimodal feedback visually, responding with subtle actions such as turning their heads to indicate attention, and visually highlighting parts of the environment to signal comprehension. These embodied behaviors, alongside visual presence, mimic aspects of human communication, allowing users to feel that the agent is engaged, present, and understanding their instructions as they speak, reinforcing a shared understanding of the task in progress.

We explore these concepts within a relatively new task context where users are given the ability to prompt a system to change a VR environment. Recent efforts (e.g. [3, 18, 71]) have demonstrated how systems can be designed to leverage large language models to create and modify objects, scenery and behaviour of objects in these environments. Manesh et al. [3] show that people's *in situ* prompts are fundamentally situated to objects in the environment, where they make reference to specific objects via gesturing and gaze. The characteristics of this task context make it suitable for our explorations into embodiment and multimodal behaviours of agents.

We conducted a Wizard of Oz study to investigate how users interact with an agent that provides non-verbal feedback while they issue verbal instructions. Participants instructed the agent to design and move objects within a VR scene, while the agent responded through multimodal behaviours and by acting on the instructions. Importantly, our agents would make mistakes 30% of the time with their multimodal signaling, and subsequent actions. We expected that the multimodal signaling cues could prevent errors, and that they provided insight into the agent's "thought process," which would thereby affect how people would repair the interaction. Our results show that with certain types of multimodal signaling cues, participants were able to prevent errors 34% of the time. We also found that when people needed to repair the interaction with the agent, they would make use of the information from the multimodal signal. Across different variations of the agent's behavior, we found that visible grounding cues, including head movements and environmental changes, were highly valued: participants expressed more confidence in the agent's understanding when visible grounding cues were present, even when errors occurred.



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713110

Our findings suggest that local grounding, enabled by non-verbal communicative multimodal behaviors such as head orientation and visual spatial cues, plays a useful social and functional role in shaping user interactions with embodied agents. This highlights the need to design agents that respond with more than just speech, integrating physical and visual acknowledgments to build a sense of shared understanding in real-time interactions.

We make two contributions in this work:

- First, we demonstrate the positive impact of local grounding behaviours in human-agent interaction;
- Second, we illustrate how these behaviors can be effectively designed for agents that interact through both head movements and environmental cues.

2 Background

2.1 Common Ground, Multimodal Signaling & Joint Attention

Effective communication requires a cooperative effort in which all parties work to both understand and be understood [1]. However, in order for this collaboration to work, there needs to be a baseline understanding between people. Common ground refers to the mutual knowledge, beliefs, and assumptions that are shared between participants of a conversation that enable effective communication [17]. This shared knowledge can take on a range of forms such as the culture in which the conversation is taking place, the cultural backgrounds of the participants, the personal histories between participants, a shared specialized set of knowledge, or information that is local to the conversation itself[16]. Each conversation participant maintains their assumptions of what knowledge is a part of the conversation's common ground. This common ground helps conversation partners to anticipate what the others might be thinking and therefore helps a conversation partner to craft their communications. However, these assumptions are not always correct, so effective communication requires the participants to actively establish, coordinate, and maintain what they believe to be common ground for the conversation in a process referred to as grounding [17].

While some of the strategies that interlocutors use to establish common ground happens in verbal conversation, much of grounding occurs in multimodal signals such as non-lexical utterances like an *mm-hm* or gestures like nodding [16]. These signals help listeners to provide feedback to an active speaker without taking their own conversational turn. A simple head nod from a listener may indicate that they understand whereas a confused vocalization may indicate that the listener needs further explanation [27]. This background feedback can help let an active speaker identify when communication has broken down and needs to be addressed. In particular, this helps the speaker initiate repair of their communication without break their conversation turn [10]. Among these communication tools, one of the earliest that we develop is joint attention. Joint attention is the use of gaze and gesture to share attention to a referent object between conversation partners [44]. For infants, joint attention is the first way they begin to understand the perspectives of others and share their perspectives and serves as a stepping stone to their development of language [43]. Even later in life, gaze [57] and referential gestures [28] serves an important role of how we make sense of our conversation partner's perspective and express our own. In this work, we explore how non-human agents may utilize the tools of multimodal signals (as manifest by joint attention) as grounding behaviours to better support understanding in human-agent interactions.

2.2 Multimodal Signals from Virtual and Physical Agents

Given the importance of multimodal signaling for effective communication, prior work has explored a variety of ways to convey such signals in avatar-mediated (including robots and virtual avatars) interactions. Among these, social gaze has been identified as an effective mechanism for improving communication quality through grounding and other mechanisms. For instance, Mutlu et al. [45] investigated the influence of a robot's gaze frequency on human-robot interactions, demonstrating that higher gaze frequency improved both participants' recall performance and their perception of the robot. In another work, they explored how robots could employ gaze cues to define and signal conversational roles, finding that participants' behaviors corresponded closely with the roles communicated through these gaze patterns[46]. Andrist et al. [5] introduced bidirectional gaze coordination for virtual characters, incorporating both the detection and production of gaze cues to foster more effective collaboration with human users during task-oriented interactions. Junko et al. [29] found that visualizing avatars' gaze targets, especially targets of joint gaze, facilitated the initiation of informal conversations in social VR.

We expand on these works to focus on gaze not only as social mechanism, but a tool in which agents may utilized to help support communication of understanding.

Another key mechanism is the use of gestures, which serves as an effective complement to verbal communication. For instance, Cassell et al. [13] and Kim et al. [31] designed virtual agents that simulate human-like interactions by integrating gestures signal changes in emotion and engagement. Narayana et al. [47] showed that an embodied agent can use gestures alone (including social, deictic, and iconic) to build enough common ground and accomplish collaborative construction tasks with human partners. In the realm of physical task support, GestureMan [32] was developed as a robot designed to improve remote instruction. This robot allows a remote instructor to control its gestures, providing clearer instructions and grounding information to supplement verbal communication. Furthermore, Rabinowitch et al. [58] investigated the impact of various gesture types, styles, and orientations on the effectiveness of collaboration between humanoid robots and humans, highlighting the nuanced ways in which gesture design influences interaction outcomes. Expanding on this body of work, we investigate how gaze and non-verbal gestures can help support grounding when the agent is acting as a listener and in particular, and investigate how these interactions can help a human to understand the state of the agent.

Previous research [41, 50, 51, 67, 69] has explored the role of multimodal signals in grounding within conversation. For instance, Marsi et al. [41] investigated the use of a talking head to visually express uncertainty through facial expressions, aiming to help users assess the reliability of information provided by the agent.

This approach contributes to more stable dialogue grounding by enabling users to gauge the agent's confidence. Pejsa et al. [50] proposed an uncertainty modeling framework that allows agents to express emotions such as bewilderment, confusion, and understanding through a combination of facial expressions and eye gaze. By coordinating verbal and nonverbal behaviors, this approach facilitates natural grounding in human-agent conversations. Perlmutter et al. [51] developed the LUCIT system, which integrates multimodal signals—including speech, pointing, and gaze—to improve grounding in human-robot interactions. By enabling the robot to use arm gestures to point at objects and confirm user intent, LUCIT enhances communication accuracy and reduces misunderstandings. Building on these findings, our study extends the exploration of multimodal signaling in grounding by focusing on how embodied agents in virtual contexts leverage gaze and spatial cues to establish a shared understanding in real-time interactions.

2.3 Embodiment

Conversational interface research has found that embodiment, often in the forms of virtual or physical 3D avatars and robots, can facilitate communication between conversational systems and humans [9, 13]. Having a body enables agents to make use of nonverbal multimodal signaling cues, including gestures, gaze, and other body movements, to establish common grounds and joint attention as discussed above. In addition to their functional values in promoting effective information exchange, embodiment enhances the social connection between agents and humans [7, 30, 33, 39]. Power et al. [52] found that robots were seen as more engaging, helpful, and lifelike in health consultation than agents displayed on flat screens. In a similar vein, Lee et al. [34] showed that people had a more positive perception of a robot dog as a companion than the same dog shown on a screen, and considered their interactions with the robot dog as of higher quality. An embodied presence could also foster trust between local and remote collaborators in remote collaboration. For example, in the study of Rae et al. [54], participants placed more trust on remote collaborators who joined through a telepresence robot than those who were displayed on a tablet.

In this work, we explored the multimodal signaling capability of a VR robotic embodiment for AI agents. Following prior work [6], our VR robotic embodiment used gaze direction to signal acknowledgement and joint attention. At the same time, by using this design, we aimed to study whether the presence of a robotic embodiment affects people's perception of the AI agent as a social actor.

2.4 Communicating Intelligent System States with Spatial Visualization

Research in human-robot teaming has recognized the value of supporting humans' situation awareness, a person's understanding of what is happening in the current situation [21], in enabling effective coordination between humans and robots. Spatial visualizations, often in the form of augmented reality (AR), provide in-context knowledge about how an intelligent system, such as a robot or a suite of smart home devices, are currently operating and plan to act. Robots can make their hidden internal states, such as battery levels and components status, visible to human teammates through visual overlays [38, 55, 56, 70]. In-situ visualizations can also show spatial constraints, including safety zones around robots [38] and detected obstacles [55]. Critical to coordination, spatial visualizations can communicate robots' intended actions so that human teammates can plan ahead and respond accordingly. Movement direction and path overlays could help people sharing the same space navigate safely around the robot [68], or choose actions that improve team efficiency [66]. AR visualizations can also show robots' intended manipulation targets [4] and trajectories [59] to better prepare human teammates for monitoring robot progress and performing complementary actions.

In our work, we use spatial visualizations to signal not only intended actions but also immediate understanding of instructions. In particular, we were interested in whether this immediate feedback could enable timely detection and correction of agent errors, in addition to the more general goal of supporting human understanding of agent intentions.

3 Design

This work focuses on the visual and interaction design of an embodied agent that supports situated prompting for a VR scene. To motivate our interaction design approach, we first define the task context, describing recent related work in the space. We then outline a set of design goals for agent multimodal signaling behaviour motivated from prior work and our own explorations. Finally, we illustrate how our final design used in the user study manifest these design goals.

3.1 Task Context: Embodied Prompting & Agent Role

The work we report on here is situated within a broader project where we are interested in "situated programming," where people can program the environment that they simultaneously inhabit. Examples of this include future visions of human-robot interaction, or enhanced versions of smart-home interaction. We explore this concept through scene/environment manipulation in a virtual reality space. Here, people can, using voice and bodily gestures, manipulate the contents of the scene, the placement and orientation of objects, and modify functionality of objects. For example, as illustrated in Figure 1, a person can change the position of object through a multi-modal prompt involving voice, gaze and gesture. The agent functions as a sort of programming agent, where it interprets the person's prompt, and translates this into functional results in the scene.

Researchers design tools to create and design VR environments and experiences with agents through prompts (e.g. [3, 18, 71]). De La Torre et al. [18] show how models can be prompted to create these experiences. Like us, Manesh et al. [3] focus on embodied prompting, where users can engage in this design creation while situated within the VR environment itself. Zhang et al. [71] explore a wider space of co-creation approaches, where the agent can provide generative suggestions through visual wireframing. In these works, while the agents nominally respond to verbal input, the interaction is far from conversational—not only in that the agents do not respond verbally, but also in how turn-taking is expressed. In each of these, a user's verbal expressions are parsed by a language model,

and then transformed into a set of programming expressions for the system to enact. In this sense, these systems are effectively voice-driven, instruction-oriented systems that allow for a broader range of inputs due to the use of a large language model. One challenge with this approach is that errors in the system's interpretation of the prompt can be corrected only after the system takes an action.

Our work departs from this prior work in two important ways: first, we embody the agent in the VR scene itself (so that users can see some visual representation of the agent); second, we imbue the agent with the ability to provide multimodal feedback to the user about how it interpreted the user's prompt. Based on common ground theory, this multimodal signaling to the user should perform two functions: first, it should give users confidence that the agent understood the user's instructions, and second, in the case of the agent misinterpreting the prompt, the user now has the opportunity to prevent the agent from making an error altogether. In our approach, we model an embodied agent that understands what the user is saying—for instance, as they refer to particular objects or locations—and expresses this understanding back to the user as the user prompts in real-time.

3.2 Design Goals for Agent Behaviour

Based on this prior work, we articulate three design goals for our embodied agent:

- *DG1. Companion rather than system.* Our goal here was to help users distinguish between the "agent" and the "system." That is, the agent should feel like a fellow collaborator that can enact changes in the system (i.e. the VR scene), and that the agent is independent of the VR scene system itself.
- *DG2. Multimodal signaling as user moves and speaks.* The agent should convey its understanding of the user's movements and prompts visually.
- *DG3. The agent should convey its state.* The agent should convey its state to the user—i.e. whether it is listening, or processing the user's inputs.

3.3 Agent Embodiment, Multimodal Signaling and Interaction Design

Avatar Companion. (DG1) In our design, we chose a floating red spherical avatar. The avatar floats at about eye-level of the user (1.5m above the floor), moves about the room, and reorients its head. Its behavioural design mimics that of a companion pet. The avatar follows the user, and stays within 0.5m-2m away (i.e. flying around as necessary), and within 30° of the forward vector of the user. The avatar will also face the user if the user is looking directly at the avatar. And, if the user is looking at a specific object in the scene, the avatar will also turn to look at the object. In both cases, "looking" is determined by the central "look vector" of the user.

The avatar is relatively small so that it does not obscure the scene. The avatar can move at a rate of 2m/s, and rotates at a rate of 1800°/s. Finally, we modeled a "processing" time for the agent, where it takes time for the agent to enact the actions that the user specifies. This was set to 4s, which was a typical lower bound from the resolver function of related work [3].

Multimodal Signaling Understanding. (DG2) As illustrated in Figure 2, we designed two methods for the agent to convey its understanding of the user's prompt. This expression was a way for the user to feel that they shared local common ground with the agent—in terms of how the agent understood user's prompt in relation to the environment and objects in it.

Figure 2: The agent expresses its understanding through multimodal signaling. (A1&A2) The agent can do this by turning its head to look at targets that are referred to by the user. (B1&B2) And by creating throbbing discs under objects or targets that are being discussed.

On-Avatar Multimodal Signaling Once it is clear to the agent what the user is referring to, the avatar cycles through looking at the objects/destinations in the scene and the user in turn. For instance, if the user were to say, "Put this object here," (as in Figure 1), the agent would cycle between looking at the object, the destination target, and the user. This is modeled after how we understand people use head orientation to interpret and understand collaborators' intentions (e.g. [22, 24, 25]).

Spatial Cues We also designed a second mechanism to express this understanding, where the agent creates circular discs underneath objects or destinations that are being referred to by the user's prompt. These circular discs turn and throb (grow and shrink in size) for visibility. This is modeled around how mixed reality interfaces and games indicate targets for the user and player to attend to [19].

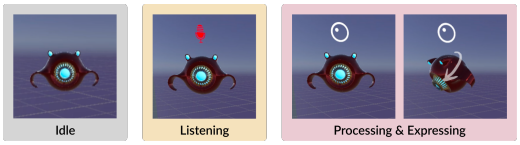


Figure 3: Behaviors and states of the agent avatar. In the Listening state, a red microphone icon signals that it is listening to users' verbal instructions. In the Processing&Expressing state, it shows a spinning icon and swivels its orientation to express its understanding of the objects and locations users have indicated.

Conveying Internal State. (DG3) The agent has four internal states, where it is "idle" and only following the user, "listening" to the user's prompt, "expressing" to the user what it understood

of the user's prompt, and "processing", where it is translating the user's prompt into instructions for the system. As illustrated in Figure 3, the listening state is denoted by a microphone icon above the avatar. "Processing&Expressing" is denoted by a spinning icon and the actual multimodal signaling behaviors (head turn and the throbbing discs). The idle state is denoted by the absence of an icon.

When the user wants the agent to "listen" them, they press and hold the left trigger of a handheld controller.

3.4 Implementation for Study

We implemented the avatar behaviour via a Wizard of Oz method, wherein the avatar's state was triggered by an experimenter. Figure 4 illustrates a flowchart of the interaction as used by the Wizard (since they were the ones implementing the behavioural design).

In a normal case, the agent begins in the idle state. When the user is done prompting the agent (i.e. they have identified all targets for that step), the agent moves to the expressing (i.e. multimodal signaling to the user about its understanding of the user's prompt) & processing state. After four seconds of processing, changes are made to the scene (nominally because of the agent), and the agent goes back to the idle state.

In some cases, we deliberately made the agent "misunderstand" the user's prompt–i.e. an error case. For instance, it might misunderstand the destination intended for a prompt such as *"Move this box over here,"* and highlight the wrong destination. In this case, after the user finishes prompting the agent, the agent expresses an erroneous signal (i.e. highlighting the wrong location). The user then has four seconds to interrupt or correct the agent's action by pressing and holding the left trigger on the handheld controller, prompting the agent to halt the incorrect action and perform the correct one. If they do not notice the erroneous signal or do not interrupt (i.e., press the left trigger within 4s), the agent moves the system to an incorrect state (e.g. moving the box to the wrong location). The user then needs to prompts the agent to move the scene back a step. Now, they can reprompt the agent. If the user does notice the erroroneous signal, and corrects the prompt, the agent displays the correct signal before making the correct change s to the scene 4s later. Such an interaction is illustrated in Figure 5.

We implemented three variations of the agent embodiment for use in the study:

Avatar-only In this variation, the spatial cues are not shown during the expressing state.

Spatial-only In this variation, the avatar is not displayed. The icons are instead attached to a transparent canvas located 1m away from the user—akin to a heads-up display in AR displays.

Full In this variation, both avatar and spatial cues are shown during the expressing state.

4 Evaluating the Impact of Agent Embodiment on Prompting Behaviour

We view situated prompting—where a user prompts a system agent with actions/commands that relate to the world that the user inhabits—to be increasingly commonplace as voice agents gain capabilities to understand their environment. Several works have begun to explore how prompting an agent that acts on the environment within a VR context (e.g.[3, 18, 71]). Manesh et al.'s focus is closest to ours here, as they study how the user prompts when they also inhabit the scene while prompting [3].

Our interaction design was focused on making it easier for people to build common ground about the local environment with an agent that they command. Based on common ground principles, we designed an embodiment for the agent (so people could direct their prompts to it), and multimodal signaling behaviours (for the agent to convey its understanding to the user to establish their local common ground).

This design led us to ask four questions that we aimed to address in our study:

- RQ1: How does the agent's multimodal signaling behaviors affect user prompting?
- RQ2: Does multimodal signaling help prevent user errors?
- RQ3: How do users repair the interaction when the agent misinterprets prompts?
- RQ4: How do users perceive the agent's embodiment?

We conducted a study where participants prompted the agent to realize various simulated environment edits in immersive VR (e.g. Figure 6). In a within-subjects design, we compared three different variations of our embodied agent against a baseline to understand how each aspect of the design affected participants' behaviours. Participants would prompt the agent to edit the scene in some way (e.g. to move an object to another location in the scene), the agent would acknowledge this through signaling, and then after some "processing time", would enact the prompt. Notably, our protocol simulated misinterpreting participant's prompts during the study procedure (30% of prompts were misinterpreted) to understand whether grounding cues would help people identify and correct AI misinterpretations. We paid attention to how participants interacted with the agent, especially focusing on how they reacted to AI misinterpretations.

4.1 Participants

We recruited 24 participants (average age: 24.75, age range: 19-31; 13 males, 11 females) through online posts and snowball sampling at a local university. The majority of them were familiar with conversational agents such as ChatGPT. Most participants had little or no experience with immersive virtual reality (16 indicated never to almost never with VR experience; only 5 indicated they used VR applications often). More information on our participants can be found in Appendix A.

4.2 Apparatus

We designed and built a custom application and VR environments for the study in Unity v2022.3.15f1. Participants performed the tasks in a 4m×4m lab space, wearing a Meta Quest 3 headset and holding handheld controllers. The experimenter used a custom web application on a laptop computer to control the VR experience.

4.3 Study Design

As our aim was to study user behaviors, we adopted the Wizard of Oz method. This was particularly important given that people prompt in different ways (e.g. [3]). Our use of the Wizard of Oz

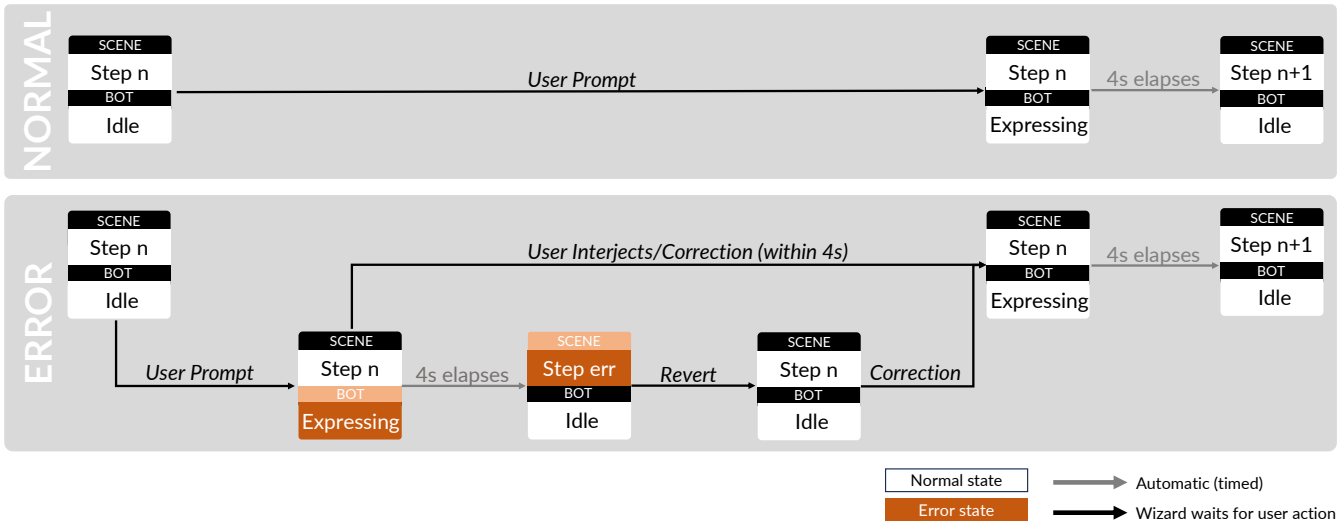


Figure 4: The interaction flow from the Wizard's perspective. (Top) In a typical case, the wizard steps after the participant's prompt, and the agent expresses multimodal signals for 4s before the scene changes to the next step. (Bottom) In an error case (specified by the protocol), after the participant's prompt, the agent expresses incorrect multimodal signals. If the participant interjects or corrects the agent within 4s of this, then the agent expresses then expresses correct multimodal signals, and the trial proceeds as normal. Otherwise, the participant needs to revert the system state, and then correct the agent with a prompt (and, this second prompting will follow the "normal" flow).

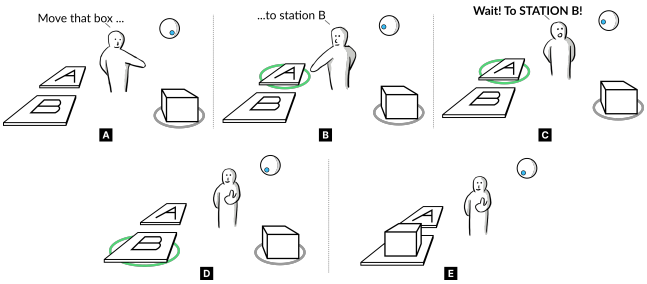


Figure 5: An example of error correction. Here, the user asks to move a box to Station B but the agent misinterprets the instruction and signals it is going to move it to Station A (*subfigure A and B*). The user catches this mistake and issues a correction (*subfigure C*). The agent communicates the corrected plan and eventually moves the box as instructed (*subfigure D and E*).

method allowed us to control for how frequently and when "misinterpretations" of the participants' prompts would occur during the study protocol. In the study, participants prompted the agent without any constraints, and an experimenter (the wizard) manually triggered the multimodal signaling behaviour, followed by predefined correct or incorrect environment changes. The study used a within-subject design where participants experienced all four conditions, with the order counterbalanced using a Latin square design.

The study included four conditions:

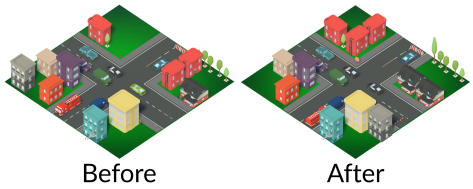


Figure 6: We built four scenes for use in the main part of the study. Above, we illustrate the "before" and "after" of Scene 4 from an isometric bird's eye view. The remaining scenes can be see in Appendix D.

Full The avatar appears along with both forms of multimodal signaling behaviour (i.e. avatar's head turns, and spatial cues appear).

Avatar-Only The avatar appears with its multimodal signaling behaviour (i.e. avatar's head turns), but no spatial cues appear.

Spatial-Only There is no avatar. The agent signals through spatial cues in the environment.

Control The agent does not present any embodiment, nor a signal.

4.4 Task Design and Study Procedure

The study began with collecting information about participants' demographics and their experience with conversational agents and virtual reality. Participants then donned a head-mounted VR display (Meta Quest 3) to complete a block of trials in the training scene.

Then, participants completed a block of trials based on the condition order specified by a balanced Latin square design. Within a block, participants experienced a VR scene. We designed a total of four scenes for the main study, in addition to a training scene. The scenes were presented in fixed order.

Within a block, participants completed 10-11 trials. In each trial, participants were introduced to the concept of a referent—a specific goal or change to be made in the scene. The referent was presented as a short 4s video they could view in world at any time to understand what action was required. The referents used in the study are detailed in Appendix D, and included tasks such as moving, rotating, replicating, or deleting objects, as well as more complex actions like manipulating multiple objects simultaneously. Participants were asked to prompt the agent to enact the goal of the referent, and could do so using voice alongside gesture and other body movements. The agent would then express its understanding through multimodal signals. After four seconds, the agent would proceed with the step that was specified by the participant. This is illustrated in Fig 4, top.

In three steps per scene, the agent would misinterpret the participant's intent (see Appendix D). As illustrated in Figure 4, in conditions involving signaling (Full, Avatar-only, Spatial-only), participants would then see an incorrect location, destination, or object highlighted. If they intervened or corrected the agent within 4s, the agent would then signal the correct prompt. If they did not intervene within four seconds, the agent would proceed with the misinterpreted action. Participants could then need to revert the state, and then correct or re-prompt the agent to achieve the correct result. This flow is illustrated in Fig 4, bottom.

After each block, participants completed a questionnaire to assess their perceptions of the agent's performance and interaction quality. The questionnaire consisted of eight items, asking participants to rate how well they felt the agent understood their prompt and responded appropriately. These items are listed in Table 1.

At the end of the experimental blocks, participants were engaged in a semi-structured interview to understand their subjective experiences with the agent. We asked participants questions that focused on their experience of the embodiment, as well as the signaling techniques. We also sought to understand how they constructed their prompts, and how the errors by the agent affected the way they conceptualized the agent.

Although participants believed they were interacting with a fully functional agent and system, it was actually controlled by a Wizard of Oz setup. At the conclusion of the study, participants were debriefed about the true nature of the system and the deception about the functionality of the agent, and given the option to withdraw their data, though none chose to do so. The protocol was approved by our Institutional Review Board.

4.5 Measures and Analysis

We recorded participants verbal utterances and other body movements through both an external camera trained on the physical test environment, as well as from the headset itself. While participants had a button to press to indicate when they expected their speech to be heard/interpreted by the agent, we recorded the entirety of their verbal utterances throughout the session. Within our application, we logged participants' head positions and orientations as a proxy for visual attention. We also logged when participants had depressed the "talk" button. From the application, we also logged program data—what participants could see in the moment, the position of those objects and their orientations, as well as study state.

We focused our coding of the experimental blocks on "error steps," where the agent nominally made errors. Specifically, we focused on what happened during these steps—did participants prevent the error? How did they try to repair the interaction with the agent? And, if they were unable to prevent the error by the agent, how did they rectify the situation with the agent?

We used Reflexive Thematic Analysis to analyze the interview data [12]. Interviews were initially transcribed using OpenAI's Whisper API[1] and then manually corrected by the second author. Next, guided by our research questions and our interest in understanding the roles of the designed components on the interaction, we coded interview transcripts without a pre-defined codebook. These initial codes were used to generate themes which guided a second round of coding.

5 Findings

We organize our findings into four subsections based on the research questions we posed at the outset. In each, we integrate findings from both the quantitative aspects of the study as well as the data from our interview components.

5.1 Perception of Multimodal Signaling

Our design focused on providing two forms of signaling: the avatar turning its head, and the spatial cues. The vast majority of our tasks focused on completing tasks where there was a source and destination (e.g. "Move the chair over there"); thus in a typical case, the agent's signaling behaviour would be to confirm the objects or locations that the participant specified (where one was an object).

Multimodal Signaling behaviour helps people feel understood. Participants discussed that signaling from the agent helped them to feel more confident in what the agent was planning on doing, as well as in their ability to formulate prompts for the agent. P4 explains that the multimodal signals are valuable, *"because I can roughly tell what will change in the scene. It helps me to anticipate what will change."* This resonates with P23's explanation that, *"I can expect what its doing right now, or what we are doing in the next steps."* Similarly, P17 describes the confidence they felt with signals, *"The highlights [the participant's description of the spatial cue signals] make the player more confident that the VR agent is understanding their words or sentence in the right way."* Participants valued feeling their prompts being understood, *"I can feel that the robot understands me... from this information, I know that 'Oh! The robot understands me.' "* Overall, 10 participants discussed that signaling from the agent made them feel more confident in interacting with the agent.

The absence of multimodal signals made participants feel anxious, and uncertain about how the agent interpreted their prompt. P24 explains, that without visuals, *"Without visual cues, ... it's not*

[1] https://platform.openai.com/docs/guides/speech-to-text

giving me any visual indication that it knows what I'm doing, and that makes me a little anxious." Similarly, P12 explains that the absence of the spatial cues, *"we don't have that sense of reassurance that it understands me. It feels a bit like there's a disconnect in terms of communication."* The signaling behaviour, whether it was via the spatial cues or the avatar's head turning behaviour, helped participants feel understood, and the absence of it was noticeable.

Spatial cues are specific and clear. Between the two mechanisms that participants used, the spatial cues and the head turning mechanisms had differential value. The spatial cues were seen as functionally useful, since they enabled participants to confirm the agent understood their prompt, and could anticipate whether it would do the right thing. Half of the participants discussed that they found the specificity of the spatial cues useful. P1 described, *"When I see the highlight, it was helpful, because it gives me the exact place."* Similarly, P2 explained that with the spatial cues, *"I can directly see if the agent got the prompt correctly or not."* P24 described the spatial cues as part of a feedback loop of communication, allowing him to validate that the agent understood the instruction, as well as for him to check what action the instruction would result in. This would allow users to, as P16, suggests, head off the action of the robot, allowing participants to predict/anticipate exactly what the agent was planning on doing.

Head turning: an ambiguous acknowledgement. On the other hand, participants found that the head turning signal was less effective at helping them anticipate whether the system understood them. For some participants, this was because the meaning of the heading turning behaviour was ambiguous for them. P14 notes that *"Because it follows your movement and it makes a lot of head movements so you can't tell which ones are meant to indicate that it's understanding of your command and which ones are just it moving."* However, even once participants understood what the head turning behaviour was indicating, they found that the head turning provided less accurate feedback into what the behaviour is referring to. P22 explains that *"I see that bot has an eye to see the objects, but sometimes you cannot precisely know what it actually sees."*. This was particularly notable when an object was clustered with other objects as P23 highlights, *"The head can only look at one direction, and that direction may contain multiple object. I don't know which one it refers to."* This overall preference towards spatial cues was also illustrated in the questionnaire data (shown in Table 1) where participants reported that they had a better understanding of whether the agent understood them and could better anticipate the agents actions in conditions with spatial cues than in conditions without.

Despite this, participants did find reassurance in the head turning behaviour and the avatar in general. While more ambiguous than the spatial cues, the head turning behaviour did still provide insight into whether the agent understood the participant. For instance, P9 felt that the head turning helped them *"know that the robot got my meaning."* Similarly, P18 explains, *"When I look at some object, the bot head also looks at some object, so I think it follows my gesture."* While the avatar did *not* follow participants' gestures, it is interesting that the turning behaviour was perceived this way. In addition, the avatar provided a sense that there was someone listening to them as they prompted the agent. We will discuss this sense of companionship that the avatar to participants provided further in Section 5.4.

In reflecting on the head turning signaling behaviour, participants pointed out that because the avatar was separated from the workspace itself, it could be hard to see the avatar, particularly what it was pointing at. P17 and P22 both explained that their focus was not on the avatar, but rather the workspace, so it was easy to miss what the avatar was doing. Instead, P17, P12 and P13 suggested that perhaps the avatar could float near the objects being talked about. P13: *"If the agent can float towards the object, and then stand right [above] it,"* this would give the clearest indication that it understood, since the reference would be to the workspace itself.

Multimodal Signaling to understand the agent's thinking process and plan. Participants found the signaling behaviour useful as it gave insight into what the agent was planning to do. These expressions of confidence in the signal are manifest in questionnaire responses. As illustrated in Table 1, participants were more confident in the agent in conditions where the agent had multimodal signal capabilities. 22/24 participants ranked either the Spatial-Only or the Full the highest when asked about their preference on conditions during the semi-structured interviews.

Nevertheless, participants wanted even more clarity into what the agent's thinking process and plans. P15 suggested that this began with ensuring that she was heard properly by the agent, suggesting that it would be valuable to see the result of the voice-to-text from her verbal instructions. This could also be at the level of how the agent would be responding to the prompt, where P22 suggested showing arrows that would explain the planned trajectory. Several participants suggested also using the audio modality for this type of confirmation. P5 explains the value of being able to see the entirety of the agent's plans, *"I want to know what the agent is thinking, so rather than be a black box, where I don't know what it is thinking,"* because then he could diagnose why it would make the mistakes that it did. Finally, P24 expected that such agents should be able to extrapolate prompts for the entire scene to multiple actions, and in this case, should pre-visualize the result of those actions for the user to see.

5.2 Error Prevention

Participants were able to use multimodal signals to prevent errors. Recall that in each block, three prompts were known to be misinterpreted incorrectly by the agent. In these cases, if the agent had signal capability, it would signal its (incorrect) interpretation of the prompt. If participants pointed out to the agent that something needed to be fixed within the 4s signal timeframe, then these would have been labeled as "error prevented" instances. Out of the 216 error trials[2], participants were able to prevent the agent error 73 times (33.8% correction rate).

How Participants Prevented Errors. We coded for how participants interjected with the agent to indicate there was an error in interpreting their prompt. We classified these based on participants' utterances, and found three types of interjections from participants:

Expressive vocalizations (E) These utterances reflect the speaker's emotional reactions or evaluations in response to the agent's signaling behaviour. These utterances are not directly part

[2] 24 participants × 3 experimental conditions × 3 error trials

Questions	Full	Avatar-Only	Spatial-Only	Control
I am able to know if the agent understood what I meant.	5.0**(0)	4.0(2.0)	5.0*(1.0)	3.0(3.0)
I can anticipate if the agent will do the right thing.	5.0**(0)	4.0*(1.0)	5.0**(1.0)	3.0(2.0)
The agent shows understanding of my prompt.	5.0*(1.0)	4.0(1.0)	5.0*(1.0)	4.0(1.0)
I am able to make sense of what the agent is doing.	5.0**(0)	4.0(2.0)	5.0**(1.0)	3.0(2.0)
It is easy to understand how the agent interprets my prompt.	5.0**(1.0)	4.0(2.0)	5.0**(1.0)	3.0(2.0)
I feel a human can understand me as well as the agent did.	5.0(1.0)	5.0(2.0)	4.0(1.0)	4.0(2.0)
The agent is good at dealing with the mistakes of my prompt.	5.0(1.0)	4.0(1.0)	5.0*(1.0)	4.0(2.0)
The agent understands well of what I am saying.	4.0(1.0)	4.0(1.0)	4.0(1.0)	4.0(1.0)

Table 1: The medians and IQRs of participants' responses to questionnaire items on a five point Likert Scale (5=strongly agree; 1=strongly disagree). * denotes that in post-hoc pairwise comparison, the responses of that condition are significantly higher than Control (p<0.05 with Bonferroni correction). ** denotes that the responses of that condition are significantly higher than both Control and Avatar-Only.

of the main communicative content, like commands. Sometimes, participants held the talk button when they said these things, other times, they did not. For instance, participants might say, "Oh!" as if they were surprised, or "Uh, no no..." (as if to themselves, or maybe the agent), or, "Oh, uh, okay...". Excerpt 1 illustrates an example of P14 reactivity vocalizing in respond to the wrong spatial cues appearing.

P14: You can move this chair to the back of the room and centralised
« Spatial cues appear »
P14: Oh no that's wrong
« Agent signaling behaviour is cancelled »
P14: Move this chair to the back over there
« Corrected signaling behaviour begins »

Excerpt 1. P14 (Scene 1, Step 4) uses an Expressive Vocalization to interject.

In some cases, participant simply vocalized their reactions without using the talk buttons and thus despite noticing the error did not cancel the erroneous behaviour as seen the following excerpt with P24.

P24: Ho ho okay
P24: Shift this table *points to the table*
P24: to touch this table *points to other table »
« Incorrect spatial cues appear »
P24: *While not holding the talk button*
Okay Boy. Wrong lah boy. Come on ah!
« Agent executes the wrong action »
P24: Undo
« Agent resets to previous state »
P24: Move that table with the books here »
« Corrected spatial cues appears »

Excerpt 2. P24 (Scene 1, Step 5) interjects with an Expressive Vocalization without prompting the agent.

Stop Commands (S) These utterances were intended as commands for the system to stop the current flow and/or to back-up to a previous step in the interaction. This is evidenced by their usually pressing and holding the talk button as they spoke. Excerpt 3 illustrates an instance of this approach, where P24 issues a command to the agent to move to the previous step.

P24: Delete that chair *points at a chair*
« Incorrect spatial cues appear »
P24: Eh? Sorry, undo
« Spatial cues stop »
P24: Delete that chair *points at the chair again*
« Corrected spatial cues appear »

Excerpt 3. P24 (Scene 1, Step 7) uses a Stop Command to prevent the agent from making an incorrect action.

In other cases, participants simply asked the agent to stop their current actions as seen in Excerpt 4 with P15.

P15: Turn this firetruck 180 degrees * points at firetruck *
« Spatial cues appear »
« Bot head begins looking at targets »
P15: Cancel prompt
« System signaling behaviour is cancelled »
P15: Turn this firetruck 180 degrees * points at firetruck again*
« Corrected signaling behaviour begins »

Excerpt 4. P15 (Scene 4, Step 3) uses a Stop Command to prevent the agent from making an incorrect action.

Interruption (I) Finally, some participants simply "interrupted" by repeating their original prompt, a revised prompt, or by providing additional information. The manner in which these corrections were constructed are the focus of the next subsection.

Table 2 shows the frequency of these interjection types that prevented an error. Each interjection type is different in whether it is meant to address the agent (Stop Commands and Interruptions address the agent, whereas Expressive Vocalizations do not necessarily address the agent), and how it addresses the agent (Expressive Vocalizations and Stop Commands are intended to signal to the

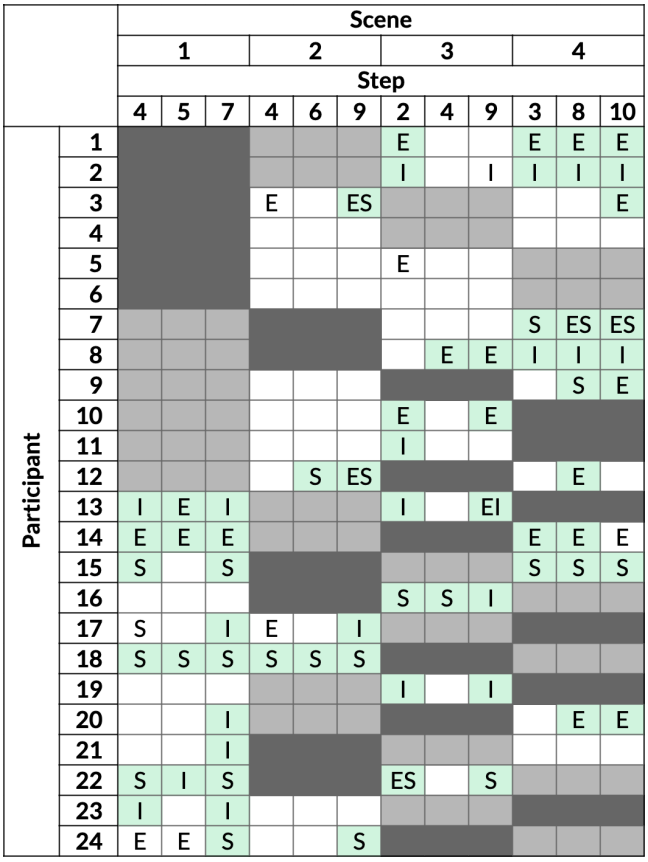


Figure 7: Interjections types per participant. (E)xpressive vocalizations, (S)top commands, and (I)nterruptions. Control blocks (no error prevention possible); Avatar-only blocks. Trials where the user prevented an error.

Expressive Vocalization	Stop Command	Interruption
25	25	23

Table 2: The count of interjections that prevented an error.

agent that something is amiss, whereas the Interruption type is essentially repeating or revising the original instruction).

Multimodal Signaling Behaviour and Error Prevention. Figure 7 provides the full coding of these interjections across all participants, as well as whether the participants were able to prevent an error. Table 3 summarizes the number of errors prevented based on condition. We see here that for most participants, they were able to use the signaling behaviours to correct the agent. A Friedman test showed a significant difference between conditions ($\chi^2 = 22.4, p < 0.001$). In blocks where the spatial cues were present (Spatial-Only and Full), participants were able to do this significantly more frequently, as shown by post-hoc Wilcoxon Rank Sum tests with Bonferroni correction ($Z = 0.0, p < 0.001$ for Avatar-Only vs Spatial-Only, and $Z = 0.0, p < 0.001$ for Avatar-Only vs Full). Consistent with participants' reflections on the value of the head-turning behaviour, we did not observe a single instance where participants were able to prevent an error in the Avatar-Only.

Condition	Errors Prevented
Avatar-Only	0
Spatial-Only	37
Full	38

Table 3: Summary of errors prevented across conditions.

This result gives strong evidence that users can use signaling behaviours to prevent errors. Yet, the design of these signaling behaviours—i.e. whether they are salient, and meaningfully linked to the effects in the environment—has an impact on this, since it is clear that in the Avatar-Only (where the backchanneling behaviour was only head turns), participants were not able to prevent any errors.

5.3 Error Correction

When the agent did make errors in interpreting the participant's prompt, we were interested in how participants attempted to repair the interaction—that is, to repair the shared, understood meaning of the prompt. We coded for how participants repaired the agent's interpretation based on participants' utterances. We observed four classes of repair utterances from participants:

Repeat (R) Participants would repeat the prompt (word for word).

Repeat with Emphasis (E) Participants would repeated the prompt, but with some emphasis on a particular word or phrase. This was sometimes accompanied by an additional gesture or bodily movement.

P3: "Remove this chair" ⇒ "Remove THIS chair, this one."

Excerpt 2. P3 (Scene 1, Step 7) repeats the prompt with emphasis.

P9: "Move the grey building to the right of the yellow building." ⇒ "Move the grey building to the RIGHT of the yellow building."

Excerpt 3. P9 (Scene 4, Step 8) repeats the prompt with emphasis.

Repeat with More information (M) Participants would repeat the prompt, but add additional information— disambiguating information.

P14: "Rotate the chair so it is facing that direction." ⇒ "Turn this chair, in front of the table, rotate it so it is facing that direction."

Excerpt 4. P14 (Scene 3, Step 2) repeats the prompt with more information.

P20: "Move the black table to next to this table." ⇒ "Move the black table, over there (pointing), to next to this table."

Excerpt 5. P20 (Scene 1, Step 5) repeats the prompt with more

Repeat	Repeat with Emphasis	Repeat with More information	Continuation
52	93	74	69

Table 4: Frequency of different repair strategies.

information and an additional gesture.

Continuation (C) Participants' utterances were a response to the agent's signaling within the context of what the participant had already said. These were usually in a short form that added additional or clarifying information.

P10: "Put this pot on the table there." ⇒ "Table, put it on the table."

Excerpt 6. P10 (Scene 3, Step 9) follows up the original prompt with a continuation that relies on the context of the interaction to interpret.

P24: "Move this car along the lane." ⇒ "The blue car, not the black car."

Excerpt 7. P24 (Scene 4, Step 10) uses a follow up that relies on the context of the interaction to interpret.

Table 4 shows a summary of the frequency of these repair utterances, while Appendix C provides a full coding of these instances.

How participants engaged in error correction suggests different ways in which participants inferred the error occurred, and their attempts to ameliorate the error. Some participants determined that in some cases, the error was because of the agent—perhaps that it had not heard them correctly, or that there was an error in the processing. In these cases, participants felt that simply repeating their prompt was sufficient, and found that this worked (Repeat strategy, 52/288 = 18.0%).

More frequently, participant felt that the source of the error was their ability to communicate their prompt clearly. For instance, P6 believed that some of the mistakes was because their prompts were not clear enough, *"Because sometimes I only use like this or that or like not a specific position. Or sometimes it mistakes between a few objects or between a few colors."*. In other cases, participants felt that it was because their gestures were inaccurate or vague. For example, P15 observed that at some points, the agent would be less accurate because they thought their hand gestures became sloppy. In these cases, participant thought that if they were more precises in both their spoken and gestured prompts, the agent would exhibit the correct behaviour. However, some participants pushed back on this notion. P5 and P24 argued that ambiguity is a natural part of communication and that dialog is important. As P5 explains,

> *"They also need a lot of explanation to understand what is going on in your head, like the ideas that you have in your head. That's why we spend so much time communicating. And like humans, there are miscommunications as well. So why would you expect a bot to be able to read your mind?"*

To resolve these ambiguities in their prompts, participants would try to clarify their intention. Repeat with emphasis was used as a strategy to emphasizing the particular idea or concept that was misunderstood, even when the rest of the prompt was understood (Emphasis strategy, 93/288 = 32.3%). The emphasis (typically a word spoken more loudly or slowly) could help an interlocutor reinterpret the prompt while focusing attention. Similarly, adding additional information (More information strategy, 74/288 = 25.7%).

The continuation strategy relied on the participant and agent having a shared understanding of the context of the interaction (Continuation strategy, 69/288 = 24.0%). In this case, the context (or local grounding) consisted of the participants' original prompt, and also the agent's signals (nominally, how it interpreted the participants' original prompt). That participants used this strategy relates to Zipf's principle of least effort, where speakers use as few words to unambiguously clarify/communicate an idea [72]. This strategy was common, appearing in over 24% of cases where participants corrected the agent.

In each of these strategies, the knowledge about how the agent interpreted their prompt, gained either through multimodal signal communication or through the agent's execution of the prompt, played an important role in how participants structured their repairs. Critically, and particularly evident in the use of the Continuation and the Emphasis strategies, the repair utterances that participants subsequently constructed show that participants expected that both they and the agent shared some local context—i.e. the previously spoken prompt, and what the agent had done so far (either show signals or completed an action). And, in making the subsequent repair statements, they aimed to repair just what had been misunderstood.

5.4 Role of Embodiment

While many participants did not find the avatar itself functionally useful, many found the "companionship" that the avatar provided reassuring. P1 suggests that the presence of the avatar, and how it behaved, *"gives me a feeling he's watching me or something, a warm feeling."* The movement of the avatar, where it follows the participant around, contributed to this feeling. P6 explained, *"I really like where the bot follows me, and follows where I'm looking at."* P7 described this feeling as being, *"I think it's more a psychological thing, like an assistant beside me, not just outside my sight."* These sentiments are echoed by other participants, who enjoyed the presence of the avatar over conditions where the avatar was not present. *"In [the experimental blocks without the avatar] it feels like I'm just talking to myself. The ones with the bot heads are a bit better."* P3 agrees, *"I feel like someone's accompanying me... I can talk with him or it."*. Overall, 9 out of 24 participants noted feeling a sense of companionship with the avatar.

The avatar thus provided a social function, where participants felt like they were communicating with an entity that would listen to them. P5 explains that, because the avatar follows him, *"I know that it is actively listening. Without the bot head, we don't know that he's listening."* Similarly, P12 describes the feeling, *"The visible agent gives a reassurance that the agent is listening. Without the the agent visible, it feels like the system is just loading, and I'm not sure whether the system is listening."* P20 describes the feeling as *"communicating with someone, but not with a virtual system."* This suggests that

its behaviour provided a sort of feeling of presence, even if the functional value of the avatar was not high in of itself.

Not all participants enjoyed the feeling of that presence. P1 likened having the avatar to being watched by a menacing creature, and would have preferred a cuter representation. P4 and P15 echoed these sentiments, but also suggested that another representation would have resolved the discomfort.

Overall, participants preferred the embodiment over the absence of an embodiment. Our questionnaire data showed that 75% (18 out of 24) participants preferred the embodiment (the Avatar-Only) compared to its absence (the Control). This was also confirmed during the semi-structured interviews where the same number of participants (18 out of 24) preferred the Avatar-only over Control. A full table describing participants' preference rankings can be found in Appendix B.

6 Discussion

Our findings provide opportunity to discuss how prompting ought to work in situated contexts—particularly if users prompt verbally. We first summarize the major findings from our study, which highlight the value of both embodiment and multimodal signaling to support users. Next, we discuss how interaction designers and researchers should consider designing multimodal signaling behaviour moving forward.

6.1 Summary of Findings

Our findings show that within the context of our study, participants were supported in their prompting by both the embodiment and the multimodal signaling behaviours of the agent. In terms of the embodiment, participants generally felt more comfortable prompting to an agent over prompting without an avatar. Participants directed their prompts toward the avatar, and looked toward it to see if the agent was listening or understood their prompts. In this sense, it functioned as a social companion.

The agent's use of spatial cues for multimodal signaling, however, were important for participants as a means to understand what the agent was planning. We saw that many participants were able to use the cues as a form of visibility into how the agent understood their prompt. This gave them confidence in their ability to prompt, and in the agent's ability to understand their prompts. Thus, the multimodal signaling was valuable to participants in feeling confident interacting with the agent.

Our findings further revealed that multimodal signaling is useful in error prevention. By leveraging the agent's multimodal signals, participants could identify potential misinterpretations in real time. User interventions to prevent errors were categorized into three types: Expressive Vocalizations, Stop Commands, and Interruptions. Among these, spatial cues were more effective in error prevention compared to head turns.

When errors occurred, participants employed different strategies to repair the interaction, including repeating their original prompts verbatim, emphasizing specific components of their prompts, providing additional clarifying information, and offering contextual continuations. These repair strategies reflected the source of the error that participants inferred. Notably, offering contextual continuations to repair highlighted that participants expected the agent has a sharing contextual understanding of the interaction. This shared context allowed participants to address specific points of misunderstanding without needing to restate their entire prompt, thereby reinforcing the interaction's efficiency and intuitiveness.

6.2 Multimodal Signaling and Models of Conversation

In our work, we applied basic principles of common ground theory, which points to the importance of the multimodal signals as a way of helping the speaker feel confident that their statements are being understood (or at least, to provide a window into how their statements are being interpreted). Building on prior work that demonstrates the value and importance of the gaze of an agent for signaling intention and interest (e.g. [13, 46]), we demonstrate that such multimodal signals are also useful to help speakers to detect breakdowns in interpretation. Furthermore, we showed that they can use these signals to build a local common ground with the agent—useful for feeling confident in cases where interaction worked well, and also useful to structure repair behaviours.

While this study was set in a virtual environment, we believe the insights could similarly inform the communication behaviors of physical robots, supporting the value of robots signaling their internal states through gaze as well as augmented-reality-based highlights [64, 66].

Our study suggests that voice-based human-agent interactions should do more than operate with a turn-based approach. Human-human conversation rarely take the form of the well structured sentences in proper turn that is common in agent-based prompting. Rather, it is common for speakers to repair their sentences mid-turn in response to both their own understanding of the common ground of the conversation, and listeners responses through the multimodal signal [17]. Even beyond the multimodal signal, listeners may interject into a speaker's turn in order to initiate repair before the utterance is completed [42].

Our design is a limited exploration of these complex repair structures in human-human conversation. We only provided a delayed signal feedback at single point to an instruction, and relied on a limited space of nonverbal communication mechanisms. Nevertheless, our findings suggests that even this limited space of multimodal communications can help participants interact more fluidly with an agent: determining whether and when a breakdown had occurred, and, they utilized strategies of emphasis and continuation which demonstrated that they assumed construction of a common ground with the agent, rather than instructing the agent from an initial state each time.

An even more effective agent responding to a user's voice instructions could provide this signal *throughout* the user's voice prompt—thereby providing a user with confidence that each part of their utterance was being understood. And, when a breakdown does occur, a user could be be more specific about what in their prompt caused the breakdown to occur. This would more closely approximate the continuous communication that human listener provides to a speaker through body language, nonverbal utterances [65] and small verbal turn breaks [62]. We demonstrated in our study that this multimodal signal can be used in cases where the

agent is confident in its interpretation, as well as (in principle) when it is uncertain (as in the error cases).

6.3 Designing Multimodal Signal Production

We consider multimodal signaling an important way for the agent to communicate its understanding of the user's intentions. Yet, how to design for more complex multimodal signaling remains a challenging design question. As our own explorations showed, even though the avatar exhibited gaze information (head turning) behaviour to participants, participants were unable to interpret the head turns in time to prevent errors. It is unclear whether this is because the avatar design was too simple, or too small, or because the avatar's behaviour was difficult to understand. And furthermore, effective design of signaling may not ultimately prevent all errors; however, as we saw, even with a modest design effort, it can help.

Another opportunity could be to consider different forms of visual spatial cues beyond the avatar. In the context of the study, participants suggested the use of arrows to depict where an object was understood to be moved, or for objects to be highlighted.

While we relied on visuals to present signaling information, we could also use other modalities. In everyday speech, for instance, we use non-lexical utterances (e.g. grunts, 'uh-huh', etc.) as ways to indicate that we understand the speaker or are uncertain of their intention [65]. An agent could use a similar kind of auditory approach to indicate whether an utterance is understood without necessarily interrupting the speaker. This approach could complement the more concrete visual ways that we present the signal right now.

Furthermore, multiple forms of multimodal signaling could support one another, where non-lexical utterances acknowledge the user's prompt, and the visual channels communicate exactly how the agent understood the prompt. Further work is needed to understand how different forms of signaling complement each other, as well as how much multimodal information is too much for users.

One limitation in our exploration of multimodal signaling is that the *tasks* that we asked participants to prompt the agent with were concrete and relatively simple. With more complex tasks, or ideas that are more abstract, it will be far more challenging to design multimodal signaling appropriately. For instance, if the prompt was more abstract (e.g. "Tidy up the room"), there may be many possible ways to fulfill the request (a five year old's mother might have a different interpretation than the five year old). Since the clean up could be fulfilled in any number of ways (and to varying levels of completeness), it is unclear whether the user needs the action to be fulfilled in a particular way, or just that the abstract form of "tidy" is sufficient. In such cases, it may be more effective to represent *uncertainty* through the signal, which would encourage the user to either continue their prompt, or to try to provide additional clarification.

6.4 Opportunities for Interaction Design

Designing for voice interaction vs conversational interaction. As designers, we need to fundamentally distinguish between an interface that supports voice interaction versus conversational interaction. Voice interaction simply means that users can issue commands and instructions verbally. On the other hand, if we are to support *conversational interaction*, this means that we need to enable the low-level communication mechanics that people use as part of conversation. Multimodal signaling is an important way that interlocutors together create and support local ground that enables both parties to have a shared understanding of the conversational state. One unique aspect of speaking is that is a performance is done a word at a time. In principle, we can interpret a user's utterances a word at a time as the words are produced, slowly building a model of the user's intention. The reflection of this model back to users is essentially multimodal signaling. If an agent is limited to audio output, then it may be suitable to express its certainty about what the user is expression (e.g. via non-lexical speech such as grunts of acknowledgement or uncertainty). On the other hand, as we illustrated in this work, it may be possible to express this understanding with other (in our case visual) modalities.

We already have at least one model of this in the form of "autocomplete" features for text-based interfaces. For instance, the search bar of popular search engines provide possible continuations of a user's query, often providing a drop-down of several possible continuations. These function as a sort of expression of the prediction agent's understanding of the user's intended query (to this point). In this case, the drop-down of possibilities is essentially asking the user to clarify between the possible continuations (i.e. by either continuing to type, or by selecting a useful continuation).

Designing for multimodal signaling with social robotics. For robotic agents in the physical world, we have already seen that such agents can use "head" orientation, gaze and gesture as part of their communicative effort [2, 6, 26]. Yet, to our knowledge, these modalities typically use gaze and gesture to communicate something to the user explicitly. It would be fruitful to explore how gaze and gesture could be used to express the agent's understanding of a user's communicative prompt, much as we have done in this work.

Yet, relying strictly on human-like ways of multimodal signaling may be unnecessarily limiting. As we showed in this work and prior research in robot design suggested [23], it is possible to use other designs that are not human-like. Equipping robots with AR and projection displays, for instance, is not entirely novel [64]; however, the point would be to use it here to express the robotic agent's understanding of the user's intentions. Such an approach would allow the user to understand the robot's plans, and the depth of its understanding of human intentions.

Multimodal signal detection by the agent: Toward Empathic agents. While this discussion has focused on the production of signaling by the agent, would could also consider agents that observe and listen for multimodal signaling behaviours from users. Such agents could use these observations to determine whether words or concepts being produced or spoken are being interpreted properly by the user, and if not, they might be able to repair the interaction without being prompted again. Such multimodal signaling might come in the form of subtle changes in facial expression, which may express uncertainty or confusion [11, 49], or other more subtle cues [40, 53]. Agents that were able to detect these forms of signaling, and use them as part of their communication with users would be perceived as more empathetic and effective at understanding the user's intentions.

6.5 Limitations and Future Works

This study employed a controlled Wizard of Oz experiment. The controlled setting allowed us to systematically compare participant interactions with the embodied agent under different multimodal signaling conditions, yielding valuable insights into the effectiveness of various design features. However, the reliance on a pre-scripted experimental protocol limited the realism and generalizability of our findings. For instance, it is not entirely natural for participants to watch a video and then recreate the virtual scene based on predefined tasks. This setup may not fully capture the complexities of real-world interactions with AI agents. Future research should employ more advanced AI models capable of dynamic, real-time interactions, enabling users to explore scenarios beyond predefined actions. Such studies would also allow for an investigation of the long-term effects of using embodied agents in naturalistic environments.

Another limitation of this study lies in the demographic composition of the participant sample. The majority of participants were university-affiliated young adults: future work should aim to diversify the participant pool to ensure generalizability of the findings. Future studies may explore the interaction between embodied agent and participants from diverse age groups, and explore how universal multimodal signaling and embodiment behavioural expectations are in respect to different cultural backgrounds, and linguistic communities.

Additionally, the relatively small sample size of this study (24 participants) may limit the robustness of the conclusions drawn. While the qualitative insights and controlled comparisons provide valuable initial evidence, future work could employ larger sample sizes to validate the observed effects and uncover potentially overlooked nuances in user-agent interactions.

7 Conclusions

This work considers how the design of a VR agent's embodiment and multimodal behaviours affects how people prompt and interact with it. Through a Wizard of Oz study, we found that participants felt more comfortable verbally prompting an embodied agent. Furthermore, we found that with appropriate multimodal signal design, users could prevent misinterpretation errors by the agent. The subsequent prompt repair behaviour that we observed showed that users expected the agent to have an understanding of their previous prompt—essentially, a shared context for the interaction. These findings can help inform the design of future AI systems involving real-time communication with users. Through proper signaling design, we can have embodied agents that are more human-like, transparent, and potentially more trustworthy. These capabilities could be particularly relevant for domains where stronger rapport matters (e.g. virtual classrooms [35], counseling [61], and performance [36]), or where effective coordination between human-agent teams underpins task success (e.g. with assistive robots [8]). They show that another way to realize the adage "reduce errors" [48] is to help the user to feel like they can build common ground with the AI agent, and that one way to do this is to allow the agent to express its understanding via multimodal signals.

Based on this work, we recommend designers and researchers consider how to design for truly *conversational interaction*, where the computing agent understands and can make use of conversational mechanics (e.g. multimodal signaling) as part of its communicative repertoire. Such an approach should help users feel that their interactions and prompts are heard and understood by the agent, and help prevent errors and surprises.

Acknowledgments

This project was funded in part by the Singapore Ministry of Education AcRF Tier 1 22-SIS-SMU-034, 22-SIS-SMU-052 and 22-SIS-SMU-092, MITACS Globalink Program, NSERC Canada Graduate Scholarships and Foreign Study Supplement, and AI SG. This research is also supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (Project ID: T2EP20220-0016). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] 2002. Logic and Conversation. In *Foundations of Cognitive Psychology: Core Readings*. The MIT Press. https://doi.org/10.7551/mitpress/3080.003.0049 arXiv:https://direct.mit.edu/book/chapter-pdf/2410650/9780262278263_cbf.pdf

[2] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

[3] Setareh Aghel Manesh, Tianyi Zhang, Yuki Onishi, Kotaro Hara, Scott Bateman, Jiannan Li, and Anthony Tang. 2024. How People Prompt Generative AI to Create Interactive VR Scenes. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 2319–2340. https://doi.org/10.1145/3643834.3661547

[4] Rasmus S Andersen, Ole Madsen, Thomas B Moeslund, and Heni Ben Amor. 2016. Projecting robot intentions into human environments. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 294–301.

[5] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2571–2582.

[6] Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3603–3612.

[7] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 701–706.

[8] Tapomayukh Bhattacharjee, Maria E Cabrera, Anat Caspi, Maya Cakmak, and Siddhartha S Srinivasa. 2019. A community-centered design framework for robot-assisted feeding systems. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 482–494.

[9] Timothy Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. *Advances in natural multimodal dialogue systems* (2005), 23–54.

[10] Leslie M. Blaha, Mitchell Abrams, Sarah A. Bibyk, Claire Bonial, Beth M. Hartzler, Christopher D. Hsu, Sangeet Khemlani, Jayde King, Robert St. Amant, J. Gregory Trafton, and Rachel Wong. 2022. Understanding Is a Process. *Frontiers in Systems Neuroscience* 16 (2022). https://doi.org/10.3389/fnsys.2022.800280

[11] Niklas Borges, Ludvig Lindblom, Ben Clarke, Anna Gander, and Robert Lowe. 2019. Classifying Confusion: Autodetection of Communicative Misunderstandings using Facial Action Units. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 401–406. https://doi.org/10.1109/ACIIW.2019.8925037

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[13] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (Dec. 2001), 67. https://doi.org/10.1609/aimag.v22i4.1593

[14] Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, Hao Yan, et al. 2000. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents* (2000), 29–63.

[15] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture &

spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[16] Eve V. Clark. 2015. *Common Ground*. John Wiley & Sons, Ltd, Chapter 15, 328–353. https://doi.org/10.1002/9781118346136.ch15 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118346136.ch15

[17] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[18] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3613904.3642579

[19] Kody R. Dillman, Terrance Tin Hoi Mok, Anthony Tang, Lora Oehlberg, and Alex Mitchell. 2018. A Visual Interaction Cue Framework from Video Game Environments for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173714

[20] Massimo Donini, Cristina Gena, and Alessandro Mazzei. 2024. Multimodal Strategies for Robot-to-Human Communication. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 417–421.

[21] Mica R Endsley. 2012. Situation awareness. *Handbook of human factors and ergonomics* (2012), 553–568.

[22] Barrett Ens, Joel Lanir, Anthony Tang, Scott Bateman, Gun Lee, Thammathip Piumsomboon, and Mark Billinghurst. 2019. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies* 131 (Nov. 2019), 81–98. https://doi.org/10.1016/j.ijhcs.2019.05.011

[23] Hadas Erel, Tzachi Shem Tov, Yoav Kessler, and Oren Zuckerman. 2019. Robots are always social: Robotic movements are automatically interpreted as social cues. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[24] Mike Fraser, Steve Benford, Jon Hindmarsh, and Christian Heath. 1999. Supporting awareness and interaction through collaborative virtual interfaces. In *Proceedings of the 12th annual ACM symposium on User interface software and technology (UIST '99)*. Association for Computing Machinery, New York, NY, USA, 27–36. https://doi.org/10.1145/320719.322580

[25] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (Sept. 2002), 411–446. https://doi.org/10.1023/A:1021271517844

[26] Zhao Han, Yifei Zhu, Albert Phan, Fernando Sandoval Garza, Amia Castro, and Tom Williams. 2023. Crossing reality: Comparing physical and virtual robot deixis. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 152–161.

[27] Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 35, 7 (2003), 1113–1142. https://doi.org/10.1016/S0378-2166(02)00190-X

[28] Jon Hindmarsh and Christian Heath. 2000. Embodied reference: A study of deixis in workplace interaction. *Journal of Pragmatics* 32, 12 (2000), 1855–1878. https://doi.org/10.1016/S0378-2166(99)00122-8

[29] Junko Ichino, Masahiro Ide, Takehito Yoshiki, Hitomi Yokoyama, Hirotoshi Asano, Hideo Miyachi, and Daisuke Okabe. 2023. How gaze visualization facilitates initiation of informal communication in 3D virtual spaces. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–32.

[30] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social cognition* 26, 2 (2008), 169–181.

[31] Hanseob Kim, Bin Han, Jieun Kim, Muhammad Firdaus Syawaludin Lubis, Gerard Jounghyun Kim, and Jae-In Hwang. 2024. Engaged and Affective Virtual Agents: Their Impact on Social Presence, Trustworthiness, and Decision-Making in the Group Discussion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[32] Hideaki Kuzuoka, Shinya Oyama, Keiichi Yamazaki, Kenji Suzuki, and Mamoru Mitsuishi. 2000. GestureMan: a mobile robot that embodies a remote instructor's actions. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 155–162. https://doi.org/10.1145/358916.358986

[33] Sonya S Kwak, Yunkyung Kim, Eunho Kim, Christine Shin, and Kwangsu Cho. 2013. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *2013 IEEE Ro-man*. IEEE, 180–185.

[34] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human-computer studies* 64, 10 (2006), 962–973.

[35] Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie A Peppler, and Karthik Ramani. 2024. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[36] Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. 2021. More kawaii than a real-person live streamer: understanding how the otaku community engages with and perceives virtual YouTubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[38] Rasmus Skovhus Lunding, Mille Skovhus Lunding, Tiare Feuchtner, Marianne Graves Petersen, Kaj Grønbæk, and Ryo Suzuki. 2024. RoboVisAR: Immersive Authoring of Condition-based AR Robot Visualisations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 462–471.

[39] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 633–644.

[40] Yingbo Ma, Yukyeong Song, Mehmet Celepkolu, Kristy Elizabeth Boyer, Eric Wiebe, Collin F. Lynch, and Maya Israel. 2024. Automatically Detecting Confusion and Conflict During Collaborative Learning Using Linguistic, Prosodic, and Facial Cues. https://doi.org/10.48550/arXiv.2401.15201 arXiv:2401.15201 [cs].

[41] Erwin Marsi and Ferdi Van Rooden. 2007. Expressing uncertainty with a talking head in a multimodal question-answering system. In *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007)*. 105–116.

[42] Ashley Micklos and Marieke Woensdregt. 2023. Cognitive and Interactive Mechanisms for Mutual Understanding in Conversation. https://doi.org/10.1093/acrefore/9780190228613.013.134

[43] Henrike Moll and Andrew N. Meltzoff. 2012. Joint Attention as the Fundamental Basis of Understanding Perspectives. In *Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience*. The MIT Press. https://doi.org/10.7551/mitpress/8841.003.0019 arXiv:https://direct.mit.edu/book/chapter-pdf/2276902/9780262303729_cdx.pdf

[44] Peter Mundy and Lisa Newell. 2007. Attention, Joint Attention, and Social Cognition. *Current Directions in Psychological Science* 16, 5 (2007), 269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x arXiv:https://doi.org/10.1111/j.1467-8721.2007.00518.x PMID: 19343102.

[45] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. 518–523. https://doi.org/10.1109/ICHR.2006.321322

[46] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, California, USA) *(HRI '09)*. Association for Computing Machinery, New York, NY, USA, 61–68. https://doi.org/10.1145/1514095.1514109

[47] Pradyumna Narayana, Nikhil Krishnaswamy, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Kyeongmin Rim, Ross Beveridge, Jaime Ruiz, James Pustejovsky, et al. 2019. Cooperating with avatars through gesture, language and action. In *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 1*. Springer, 272–293.

[48] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[49] Mariya Pachman, Amaël Arguel, Lori Lockyer, Gregor Kennedy, and Jason Lodge. 2016. Eye tracking and early detection of confusion in digital learning environments: Proof of concept. *Australasian Journal of Educational Technology* 32, 6 (Dec. 2016). https://doi.org/10.14742/ajet.3060 Number: 6.

[50] Tomislav Pejsa, Dan Bohus, Michael F Cohen, Chit W Saw, James Mahoney, and Eric Horvitz. 2014. Natural communication about uncertainties in situated interaction. In *Proceedings of the 16th international conference on multimodal interaction*. 283–290.

[51] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency.. In *Robotics: Science and Systems*, Vol. 12. 40–50.

[52] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 145–152.

[53] Meera Radhakrishnan, Thivya Kandappu, Manoj Gulati, and Archan Misra. 2023. Wearables for In-Situ Monitoring of Cognitive States: Challenges and Opportunities. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 671–676. https://doi.org/10.1109/PerComWorkshops56833.2023.10150270 ISSN: 2766-8576.

[54] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. In-body experiences: embodiment, control, and trust in robot-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1921–1930.

[55] Christopher Reardon, Kevin Lee, John G Rogers, and Jonathan Fink. 2019. Communicating via augmented reality for human-robot teaming in field environments. In *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 94–101.

[56] Patrick Renner, Florian Lier, Felix Friese, Thies Pfeiffer, and Sven Wachsmuth. 2018. Wysiwicd: What you see is what i can do. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 382–382.

[57] Daniel C. Richardson, Rick Dale, and John M. Tomlinson. 2009. Conversation, Gaze Coordination, and Beliefs About Visual Context. *Cognitive Science* 33, 8 (2009), 1468–1482. https://doi.org/10.1111/j.1551-6709.2009.01057.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2009.01057.x

[58] Laurel D. Riek, Tal-Chen Rabinowitch, Paul Bremner, Anthony G. Pipe, Mike Fraser, and Peter Robinson. 2010. Cooperative gestures: Effective signaling for humanoid robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 61–68. https://doi.org/10.1109/HRI.2010.5453266

[59] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. 2020. Communicating robot arm motion intent through mixed reality head-mounted displays. In *Robotics research: The 18th international symposium ISRR*. Springer, 301–316.

[60] Federico Rossano. 2012. *Gaze in Conversation*. John Wiley & Sons, Ltd, Chapter 15, 308–329. https://doi.org/10.1002/9781118325001.ch15 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118325001.ch15

[61] Eriko Sakurai, Kentarou Kurashige, Setsuo Tsuruta, Yoshitaka Sakurai, Rainer Knauf, Ernesto Damiani, Andrea Kutics, and Fulvio Frati. 2020. Embodiment matters: toward culture-specific robotized counselling. *Journal of Reliable Intelligent Environments* 6 (2020), 129–139.

[62] EA Schegloff. 2000. When 'others' initiate repair. *Applied Linguistics* 21, 2 (06 2000), 205–243. https://doi.org/10.1093/applin/21.2.205 arXiv:https://academic.oup.com/applij/article-pdf/21/2/205/351026/210205.pdf

[63] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4 (May 2021), 81:1–81:43. https://doi.org/10.1145/3386867

[64] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. 2022. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–33.

[65] Jacqueline Urakami and Katie Seaborn. 2023. Nonverbal Cues in Human–Robot Interaction: A Communication Studies Perspective. *J. Hum.-Robot Interact.* 12, 2, Article 22 (mar 2023), 21 pages. https://doi.org/10.1145/3570169

[66] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. 2018. Communicating robot motion intent with augmented reality. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 316–324.

[67] Chao Wang, Anna Belardinelli, Stephan Hasler, Theodoros Stouraitis, Daniel Tanneberg, and Michael Gienger. 2023. Explainable human-robot training and cooperation with augmented reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–5.

[68] Atsushi Watanabe, Tetsushi Ikeda, Yoichi Morales, Kazuhiko Shinozawa, Takahiro Miyashita, and Norihiro Hagita. 2015. Communicating robotic navigational intentions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5763–5769.

[69] Tom Williams, Matthew Bussing, Sebastian Cabrol, Elizabeth Boyle, and Nhan Tran. 2019. Mixed reality deictic gesture for multi-modal robot communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 191–201.

[70] James E Young, Min Xin, and Ehud Sharlin. 2007. Robot expressionism through cartooning. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 309–316.

[71] Lei Zhang, Jin Pan, Jacob Gettig, Steve Oney, and Anhong Guo. 2024. VRCopilot: Authoring 3D Layouts with Generative AI Models in VR. https://doi.org/10.1145/3654777.3676451 arXiv:2408.09382 [cs].

[72] George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

A Participants

We recruited 24 participants. Detailed information about the participants is found in Table 5.

ID	Gender	Age	Previous VR Experience	Condition Order
P1	M	25	Almost never	Control · Avatar-Only · Spatial-Only · Full
P2	M	21	Almost never	Control · Avatar-Only · Full · Spatial-Only
P3	M	26	Almost never	Control · Full · Avatar-Only · Spatial-Only
P4	M	21	Almost never	Control · Spatial-Only · Avatar-Only · Full
P5	F	24	Often	Control · Spatial-Only · Full · Avatar-Only
P6	F	26	Never	Control · Full · Spatial-Only · Avatar-Only
P7	F	24	Never	Avatar-Only · Control · Spatial-Only · Full
P8	M	23	Never	Avatar-Only · Control · Full · Spatial-Only
P9	F	21	Often	Avatar-Only · Full · Control · Spatial-Only
P10	M	24	Often	Avatar-Only · Full · Spatial-Only · Control
P11	F	20	Sometimes	Avatar-Only · Spatial-Only · Full · Control
P12	M	23	Never	Avatar-Only · Spatial-Only · Control · Full
P13	M	24	Never	Spatial-Only · Avatar-Only · Full · Control
P14	F	19	Almost never	Spatial-Only · Avatar-Only · Control · Full
P15	M	23	Almost never	Spatial-Only · Control · Avatar-Only · Full
P16	F	21	Sometimes	Spatial-Only · Control · Full · Avatar-Only
P17	M	26	Almost never	Spatial-Only · Full · Avatar-Only · Control
P18	F	28	Almost never	Spatial-Only · Full · Control · Avatar-Only
P19	F	26	Never	Full · Avatar-Only · Spatial-Only · Control
P20	M	31	Almost never	Full · Avatar-Only · Control · Spatial-Only
P21	F	29	Never	Full · Control · Avatar-Only · Spatial-Only
P22	F	22	Often	Full · Control · Spatial-Only · Avatar-Only
P23	M	29	Sometimes	Full · Spatial-Only · Avatar-Only · Control
P24	M	28	Often	Full · Spatial-Only · Control · Avatar-Only

Table 5: Participant data and condition order.

B Condition Preference Ranking

We asked participants to rank their preference in study conditions. This ranking is shown in Table 6

C Repair Coding

We coded how participants repaired the interaction when the agent made or expressed an error in its understanding of the participant's prompt. This coding is show in Figure 8.

	Condition Preferance Ranking (Left To Right, Most Preferred to Least)			
P1	Full	Spatial-Only	Avatar-Only	Control
P2	Spatial-Only	Full	Avatar-Only	Control
P3	Full	Spatial-Only	Avatar-Only	Control
P4	Spatial-Only	Full	Avatar-Only	Control
P5	Full	Spatial-Only	Avatar-Only	Control
P6	Full	Spatial-Only	Avatar-Only	Control
P7	Full	Spatial-Only	Avatar-Only	Control
P8	Spatial-Only	Full	Control	Avatar-Only
P9	Spatial-Only	Full	Avatar-Only	Control
P10	Both	Spatial-Only or Avatar-Only		Control
P11	None	Avatar-Only	Spatial-Only	Full
P12	Full	Spatial-Only	Avatar-Only	Control
P13	Full	Spatial-Only	Control	Avatar-Only
P14	Spatial-Only	Full	Control	Avatar-Only
P15	Spatial-Only	Full	Avatar-Only	Control
P16	Full	Spatial-Only	Avatar-Only	Control
P17	Spatial Conditions		Non-Spatial Conditions	
P18	Full	Spatial-Only	Avatar-Only	Control
P19	Control	All Other Conditions		
P20	Full	Spatial-Only	Avatar-Only	Control
P21	Spatial-Only	Full	Avatar-Only	Control
P22	Spatial-Only	Full	Avatar-Only	Control
P23	Full	Spatial-Only	Avatar-Only	Control
P24	Full	Spatial-Only	Avatar-Only	Control

Table 6: Participant's Ranking of Conditions by Overall Preference.

		Scene											
		1			2			3			4		
		Step											
		4	5	7	4	6	9	2	4	9	3	8	10
Participant	1	M	R	M	M	E	E	R	R	M	C	C	C
	2	C	R	M	R	R	C	C	R	C	C	C	C
	3	R	R	E	E	C	E	E	R	E	M	M	E
	4	C	M	M	E	E	C	C	R	C	C	C	C
	5	E	C	M	M	M	M	M	M	C	C	E	E
	6	M	M	M	E	R	R	E	E	E	R	R	E
	7	C	C	C	R	C	E	M	M	M	E	M	E
	8	R	R	E	E	R	R	E	E	C	C	C	C
	9	M	M	E	R	C	E	R	E	E	E	E	M
	10	E	M	C	C	C	C	C	C	C	C	C	C
	11	E	M	M	E	R	C	M	C	C	R	C	E
	12	E	E	E	M	R	E	C	R	C	C	C	C
	13	M	E	M	E	E	C	M	M	C	E	E	M
	14	R	R	M	M	C	M	M	E	C	R	M	C
	15	R	E	E	E	M	E	E	M	M	E	M	M
	16	E	M	M	E	R	E	M	M	C	C	C	M
	17	E	M	C	R	E	C	E	E	E	R	M	E
	18	M	E	M	M	R	E	E	M	E	M	R	E
	19	R	E	M	E	R	E	E	E	E	R	M	R
	20	E	M	E	R	M	M	C	R	R	E	M	E
	21	R	M	C	E	R	C	M	E	C	E	R	E
	22	R	R	M	R	E	E	M	M	R	E	M	E
	23	E	C	E	C	R	E	M	M	M	R	M	E
	24	E	M	E	E	E	R	E	E	C	C	C	C

Figure 8: Repair utterances per participant. (R)epeat ; Repeat with (E)mphasis ; Repeat with (M)ore information ; and (C)ontinuation .

D Referents in Elicitation Study

We created a total of 48 referents across 5 different scenes (including the training scene) for the elicitation study. Each scene consists of 10-11 steps, which includes tasks like moving, creating, modifying and removing objects. There are also three steps per scene where the system misinterpret the participant's intent. These are described in Table 7 and Table 8.

E Referents Used in the Study

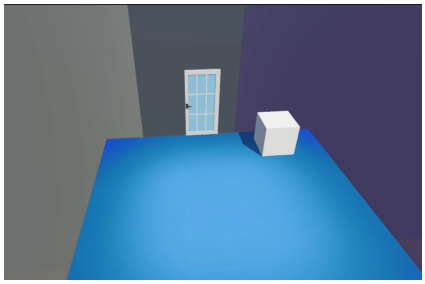
While these referents were presented to participants as short video clips (typically 4s), we reproduce the "final" frame of these referent videos to show how the participants perceived their task. For clarity, we have changed the floor texture in the reference images.

Table 7: The referents that were used in the study for Scenes 0, 1 and 2. We include here the scene number, along with the steps for each scene. Each step was its own referent.

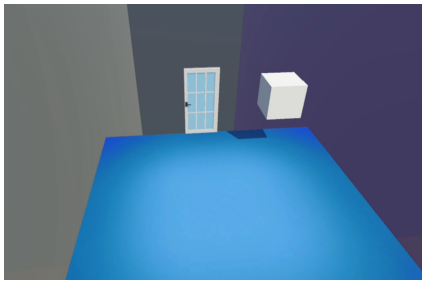
Scene (#)	Step	Task Type	Notes
Training Scene (0)	0 The box shifts from the left side of the door to the right side.	Move	
	1 The box levitates about 1m above the ground.	Move	
	2 A sphere is created in front of the box.	State Change	
	3 The sphere becomes twice its size.	State Change	
	4 The sphere is copied and placed to the left side of the room.	Copy	
	5 The sphere has a black and white checkered pattern.	State Change	
Living Room (1)	0 The most front chair is aligned with the table.	Move	
	1 The arrangement is copied to the right side of the room.	Copy Multiple	
	2 The colour of the left table changes to white.	State Change	
	3 The two top right books move to the right table.	Move Multiple	
	4 *The chair in the middle is moved out of the way.*	Move	The chair is moved to the wrong side of the room
	5 *The black table is moved closer to the white table.*	Move	The wrong table is moved
	6 The leftmost chair changes to blue.	State Change	
	7 *The middle chair is removed.*	Remove	The wrong chair is removed
	8 The two right chairs are aligned with the black table.	Move Multiple	
	9 The white table changes back to black.	State Change	
Bedroom (2)	0 The teddy bear is copied 2 times.	Copy	
	1 The new teddy bears are changed into a penguin and a pig.	State Change Multiple	
	2 The window on the right is placed on top of the dolls.	Move	
	3 The bed shifts to the middle of the window.	Move	
	4 *The bed is rotated ninety degrees anti-clockwise.*	State Change	The bureau is rotated instead of the bed
	5 The nightstand and the desk moves to the right of the bed.	Move Multiple	
	6 *The nightstand and desk are copied and put on the left side of the bed.*	Copy Multiple	The copy is placed on the wrong location
	7 The blanket is changed to beige colour from its original blue colour.	State Change	
	8 The tall light is removed from the corner of the room where it had been since the start.	Remove	
	9 *The pig doll is moved to the couch.*	Move	The pig is placed on the bed instead of the couch
	10 The penguin doll is moved to the bed.	Move	

Table 8: The referents that were used in the study for Scenes 3 and 4.

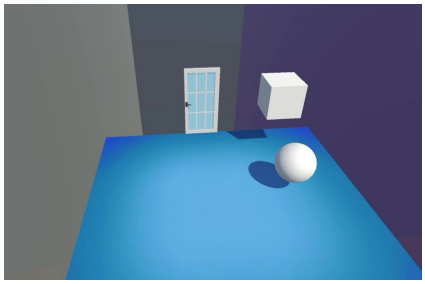
Scene (#)	Step	Task Type	Notes
Study Room (3)	0 The cupboard is shifted to the right of the lamp.	Move	
	1 The chair is copied and placed in front of the table.	Copy	
	2 *The new chair is rotated ninety degrees anti-clockwise.*	State Change	The wrong chair is rotated
	3 The lamps are turned on.	State Change Multiple	
	4 *The red lamp is turned off.*	State Change	The white light is turned off instead of the red light
	5 The books on the table are moved to the bookshelf.	Move Multiple	
	6 The book on the floor is moved to the bookshelf.	Move	
	7 The red chairs are changed to blue.	State Change Multiple	
	8 The red lamp is removed.	Remove	
	9 *The plant pot is moved to the table.*	Move	The flowerpot is placed on the chair instead of the table
Toy Cars (4)	0 The fire truck is shifted to the left lane.	Move	
	1 The two red buildings are moved to the other side of the street.	Move Multiple	
	2 The house is copied and placed on the left side.	Copy	
	3 *The fire truck is rotated to face the right direction.*	State Change	The blue truck near the fire truck is rotated instead of the fire truck
	4 The hotdog stand is moved to the opposite side of the road.	Move	
	5 The cone is duplicated four times.	Copy	
	6 The green car is changed to black.	State Change	
	7 The garbage bag is removed.	Remove	
	8 *The grey building is moved to the side of the yellow building.*	Move	The black and white building is wrongly placed on the road
	9 The most left tree becomes taller.	State Change	
	10 *The blue car is moved forward.*	Move	The black car is moved instead of the blue car



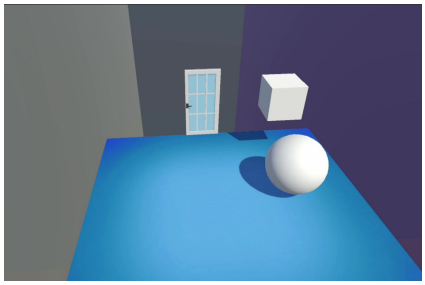
(a) Step 0: The box shifts from the left side of the door to the right side.



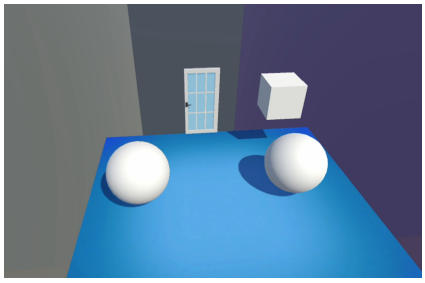
(b) Step 1: The box levitates about 1m above the ground.



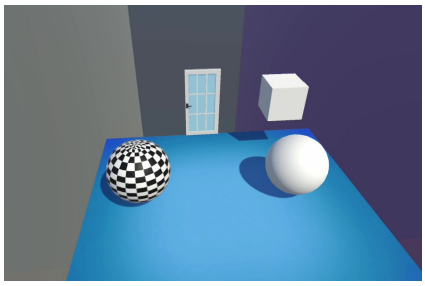
(c) Step 2: A sphere is created in front of the box.



(d) Step 3: The sphere becomes twice its size.

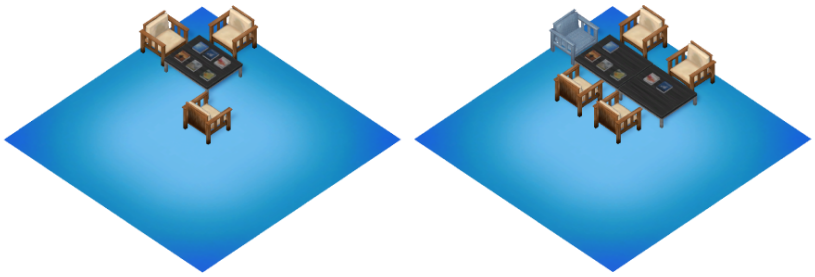


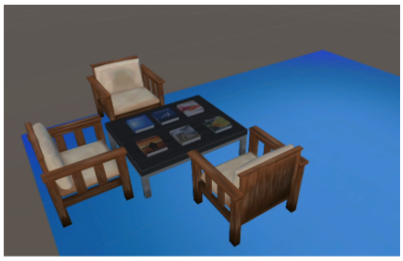
(e) Step 4: The sphere copied and placed to the left side of the room.



(f) Step 5: The sphere has a black and white checkered pattern.

Figure 9: Referent images of each step in the training scene (Scene 0).



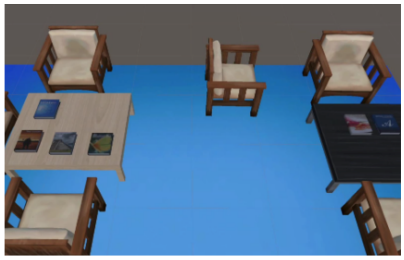


(a) Step 0: The most front chair is aligned with the table.

(b) Step 1: The arrangement is copied to the right side of the room.

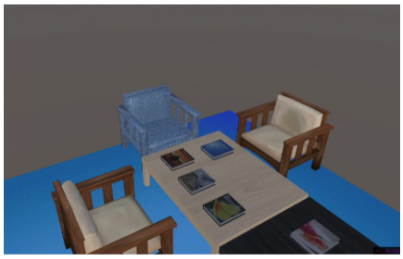
(c) Step 2: The colour of the left table changes to white.

(d) Step 3: The two top right books move to the right table.



(e) Step 4: The chair in the middle is moved out of the way.

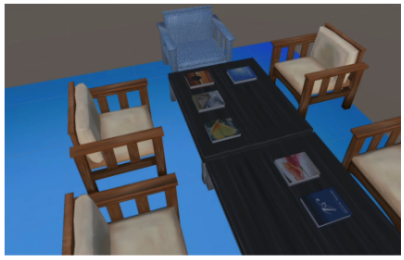
(f) Step 5: The black table is moved closer to the white table.



(g) Step 6: The leftmost chair changes to blue.

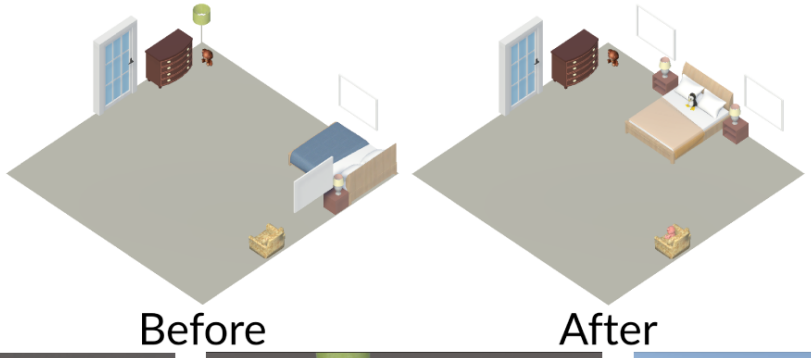
(h) Step 7: The middle chair is removed.

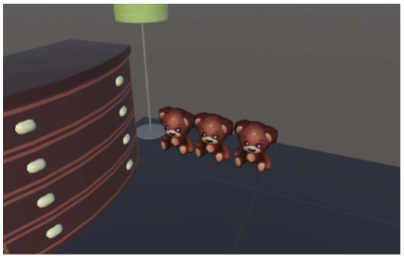
(i) Step 8: The two right chairs are aligned with the black table.



(j) Step 9: The white table changes back to black.

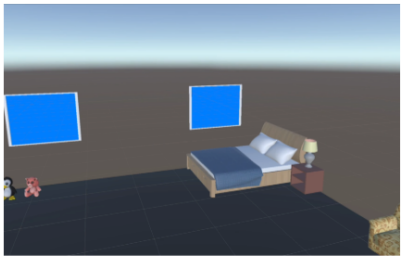
Figure 10: Referent images of each step in the Living Room scene (Scene 1).



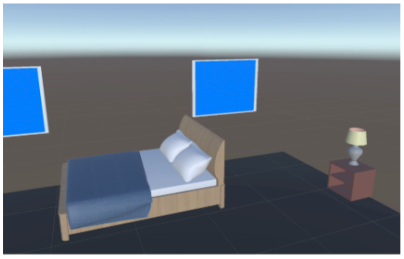


(a) Step 0: The teddy bear is copied 2 times.

(b) Step 1: The new teddy bears are changed into a penguin and a pig.

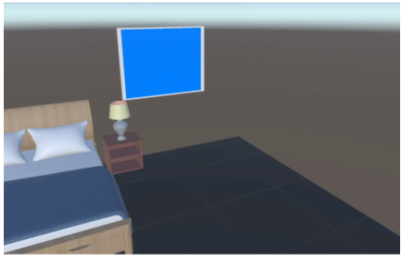


(c) Step 2: The window on the right is placed on top of the dolls.

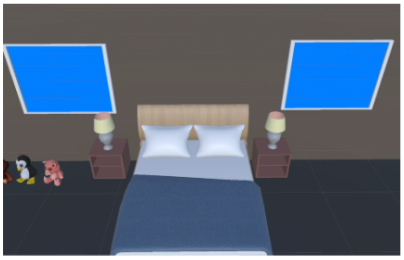


(d) Step 3: The bed shifts to the middle of the window.

(e) Step 4: The bed is rotated ninety degrees anti-clockwise.



(f) Step 5: The nightstand and the desk moves to the right of the bed.



(g) Step 6: The nightstand and desk are copied and put on the left side of the bed.

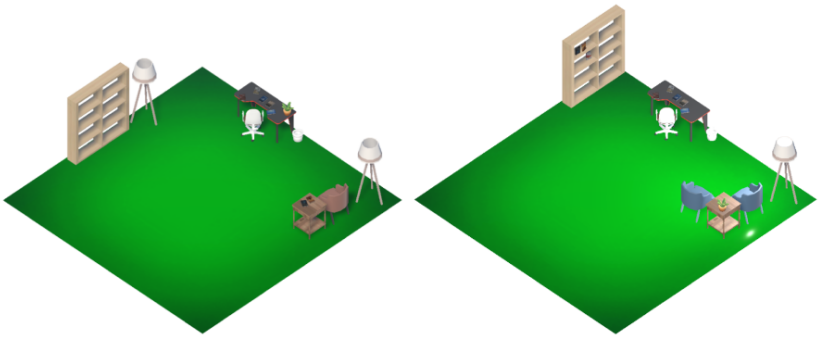
(h) Step 7: The blanket is changed to beige colour from its original blue colour.

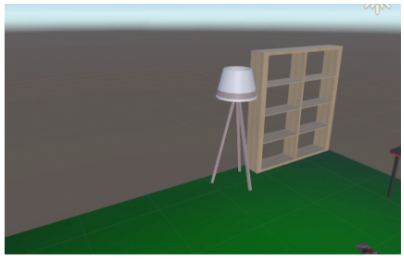
(i) Step 8: The tall light is removed from the corner of the room where it had been since the start.

(j) Step 9: The pig doll is moved to the couch.

(k) Step 10: The penguin doll is moved to the bed.

Figure 11: Referent images of each step in the Bedroom scene (Scene 2).

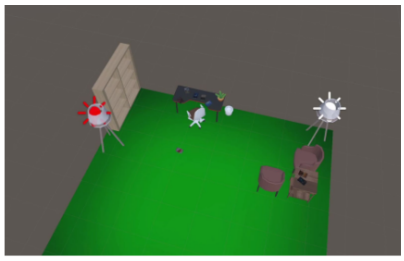




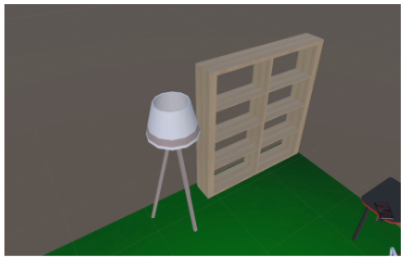
(a) Step 0: The cupboard is shifted to the right of the lamp.

(b) Step 1: The chair is copied and placed in front of the table.

(c) Step 2: The new chair is rotated ninety degrees anti-clockwise.



(d) Step 3: The lamps are turned on.



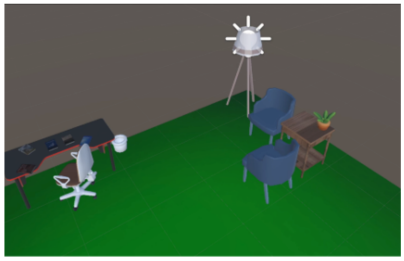
(e) Step 4: The red lamp is turned off.

(f) Step 5: The books on the table are moved to the bookshelf.

(g) Step 6: The book on the floor is moved to the bookshelf.

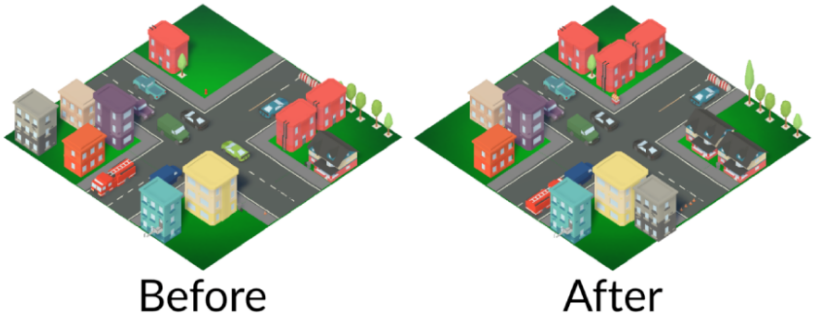
(h) Step 7: The red chairs are changed to blue.

(i) Step 8: The red lamp is removed.

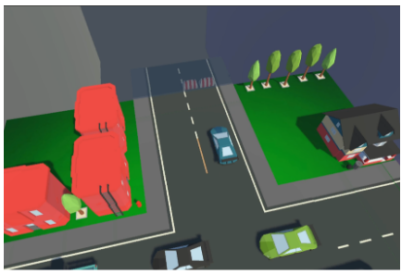


(j) Step 9: The plant pot is moved to the table.

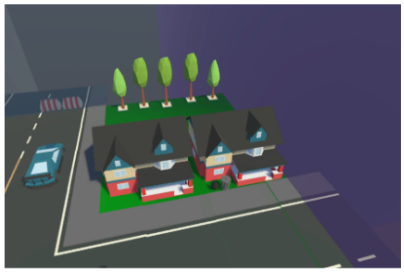
Figure 12: Referent images of each step in the Study Room scene (Scene 3).



(a) Step 0: The fire truck is shifted to the left lane.



(b) Step 1: The two red buildings are moved to the other side of the street.



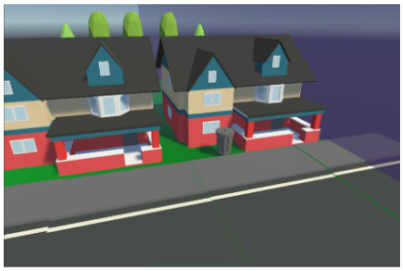
(c) Step 2: The house is copied and placed on the left side.

(d) Step 3: The fire truck is rotated to face the right direction.

(e) Step 4: The hotdog stand is moved to the opposite side of the road.

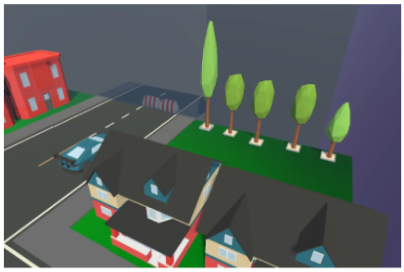
(f) Step 5: The cone is duplicated four times.

(g) Step 6: The green car is changed to black.



(h) Step 7: The garbage bag is removed.

(i) Step 8: The grey building is moved to the side of the yellow building.



(j) Step 9: The most left tree becomes taller.

(k) Step 10: The blue car is moved forward.

Figure 13: Referent images of each step in the Toy Cars scene (Scene 4).